

The proteome isn't genome II

The Human Genome Organisation was both a cheer-leader and a coordinator for genomics. But proteomics is a different beast, and the fledgling Human Proteome Organisation will struggle to find a similar role.

The Human Genome Project (HGP) had one clear goal: producing a complete DNA sequence for each of our chromosomes. This has already been attained in draft form, and a final sequence is due by 2003.

Some credit for these achievements goes to the Human Genome Organisation (HUGO), which fostered the exchange of data and materials, and helped to coordinate the activities of national genome initiatives. Although questions have been raised about its effectiveness in recent years, HUGO raised the profile of genomics and helped shape policy, including the HGP's 1996 decision that sequence data should be placed in public databases within 24 hours of collection. When HUGO was created in 1989, money was scarce and industry sceptical. The scattered community of gene mappers needed a voice.

But the HGP is also a means to an end — that being a deeper understanding of fundamental biology and disease mechanisms. This is where proteomics — the large-scale identification of proteins — is starting to come into play. And where in 1989 there was HUGO, now there is a fledgling Human Proteome Organisation, or HUPO.

HUPO will struggle to emulate its predecessor, however, because human proteomics is not a single project with one endpoint that lends itself to HUGO-style coordination. Traditionally, a proteome has been defined as the complete set of proteins that is produced by the genome during the lifetime of an organism. That is a lot of proteins — as many as half a million. And in biological terms, what really matters are the snapshots of proteins produced at a particular time, under particular conditions, by particular types of cells. Defined in this way, the human proteome is almost infinitely dynamic.

There is also little need for HUPO to copy HUGO's early cheer-leading role. Relatively speaking, money abounds. Countries that missed making a major mark in genomics — such as South Korea —

are determined to make up for it by giving support to proteomics. Investment in the private sector is also energetic.

The early involvement of industry will also reduce the scope for HUPO to assume a coordinating role. After it entered the DNA sequencing fray in 1998, Celera Genomics of Rockville, Maryland, showed no inclination to be coordinated by HUGO, and rejected the HGP's policies on data access. Celera is again among the major players in proteomics, and there is no reason to assume that its competitors will be any more willing to take directions from HUPO.

Celera and companies such as Oxford GlycoSciences and Large Scale Biology, which operates a proteomics subsidiary in Germantown, Maryland, aim to compare healthy and diseased tissues, studying tens of thousands of proteins. They hope to exploit observed differences in the number and levels of expressed proteins to develop diagnostic tests and to discover new drug targets. Many academic groups are also exploring similar avenues.

Other companies and academic groups intend to study protein-protein interactions, while yet others are determining protein structures on a large scale. Purists might argue that these projects stretch the definition of proteomics too far — but even if they are excluded, proteomics will still require a wide variety of technologies, including gel electrophoresis, to separate proteins, and mass spectrometry, to characterize them. But tackling the full range of human proteins will require techniques still under development, many of them proprietary.

HUPO's future was debated last week at meetings in York, England, and McLean, Virginia, but no clear answers emerged. Perhaps the best HUPO can do is to become a clearing-house for information on a diverse and fast-moving field, and the technologies that underpin it — a modest role, certainly, but a perfectly respectable one. ■

An absence of angst

German science seems surprisingly comfortable with the concept of a research prize sponsored by a tobacco giant.

Soul-searching is something of a national pastime in Germany — this is, after all, the culture that invented angst. It is surprising, therefore, that the German arm of the tobacco giant Philip Morris's philanthropic foundation has been awarding annual research prizes for nearly two decades without attracting any significant controversy. All the more so when you consider the furore that has erupted in Britain over instances in which academic institutions have accepted money from the tobacco industry.

Since 1983, around 100 researchers in Germany, Austria and Switzerland have received the Philip Morris prize. It is a prestigious award, given only to scientists of the highest calibre, and can be awarded in all areas of science, except medicine. This year's winners, announced last week, each take home around US\$100,000.

Tobacco firms may attract widespread opprobrium, but it can be argued that in this regard they stand out from other companies only by degree. Some high-minded individuals, concerned about drugs

pricing policy in the developing world, might refuse to accept an award from a pharmaceuticals giant; many environmentalists, meanwhile, would draw the line at accepting a prize from a notoriously polluting chemicals firm.

But over the past few years, it has become clear that there is a special irony to a company such as Philip Morris presenting a prize for outstanding science. A detailed report published last year by the World Health Organization (see *Nature* 406, 547; 2000) revealed how Philip Morris and other tobacco firms tried to undermine science-based assessments of the health risks posed by tobacco.

It could be argued that Philip Morris should be encouraged to reward excellence in research, rather than trying to manipulate science to its own ends. But against that view must be balanced the knowledge that accepting a prize from Philip Morris provides favourable publicity for a company that appears in the past to have stooped to such tactics. It is a dilemma that merits some good German soul-searching. ■





Rocky problem
Dispute flares up over access to Ethiopian fossil sites
p728



Scope for doubt
Astronomers in limbo over funding for telescope
p729



Causal link
South Africa accepts link between HIV and AIDS
p730



In the red
NASA has another crack at probing Mars
p733

Museum visitors take priority as Smithsonian curbs research

Josette Chen

The Smithsonian Institution in Washington DC, the world's largest museum complex, plans to close at least two of its research centres, with the loss of 350 jobs. Some of the museum's scientists say its administrators are turning their backs on its research mission.

Staff at the Conservation and Research Center in Front Royal, Virginia, a conservation biology facility employing 80 people, learned last week that it is to close by the end of the year. The centre conducts research on endangered species and breeding programmes under the auspices of the National Zoo, which is part of the Smithsonian complex.

Also to close is the Smithsonian Center for Materials Research and Education, which employs a smaller number of researchers developing new techniques for preserving museum artefacts.

Scientists fear that the Smithsonian will continue to cut its research activities. According to one senior researcher, who recently left the institution and wished not to be named, the announcement is "the first shoe of probably many ... this is a centipede".

The changes are "shocking and frighten-



Mammoth attraction: the Smithsonian Institution's funding sources tend to prefer popular exhibits to research.

ing for science", says Brian Huber, a palaeobiologist at the National Museum of Natural History and chair of the Smithsonian's academic senate. The senate will meet on 16 April to discuss the plans, but has no power to reverse them. They are likely to be confirmed by the Smithsonian's Board of Regents when it meets in May.

The Smithsonian Institution was set up

in 1846 with funds bequeathed by James Smithson, an English chemist and geologist, to promote "the increase and diffusion of knowledge". But its annual operating budget of almost \$600 million comes from the US Congress and from private donors — both of which tend to attach more importance to popular exhibits than to research.

Researchers at the museum complex believe that Larry Small, a former banker who was appointed last year as the first non-academic to serve as the Smithsonian's secretary, is behind the changes. "I feel strongly that Larry Small is not the right person to be in charge of the Smithsonian," says David Swofford, an evolutionary biologist who is leaving the natural-history museum for a post at Florida State University. "Morale is lower than it has ever been."

Dennis O'Connor, undersecretary of science at the complex, declined to comment on the closures, which are part of the 2002 federal budget proposal that was announced on Monday (see page 731). The overall budget for the museum complex is set to grow modestly in 2002, however.

A memo issued by Lucy Spelman, director of the National Zoo, says that the conservation centre's closure is a result of budget constraints. But the memo twice mentions that funds released by the closure will be redirected to support public services at the Smithsonian.

UK foot-and-mouth epidemic slows

Jim Giles, London

The foot-and-mouth outbreak in Britain might at last have peaked, say the government's scientific advisers.

Epidemiologists working with the Ministry of Agriculture, Fisheries and Food (MAFF) say that a levelling off in the number of new cases being reported every day indicates that the disease may now be coming under control.

Last month, three independent groups of researchers told the government that rapid slaughter of infected animals and those on neighbouring farms was the best way to slow the outbreak (see *Nature* 410, 501; 2001). MAFF says that infected animals are now being slaughtered about 36 hours after they are identified, compared with the previous delay of two or three days.

"We are much more confident now than we were last week that it has peaked," says Neil Ferguson, an epidemiologist at Imperial College, London, who is advising MAFF.

But the precautionary cull might also be artificially lowering the number of new cases. Some new outbreaks will be unreported as the animals are killed before they are diagnosed. The epidemiologists working with MAFF say that the picture will be clearer by the beginning of next week.

Hopes for a speedy end to the outbreak receded last weekend, when new cases were identified in two previously unaffected areas. Government vets confirmed that the virus has spread into South Wales from infected areas in the north. The disease has also spread east and west from a pocket of infection in southern Scotland.

Restrictions delay fossil hunts in Ethiopia

Rex Dalton, Kansas City

Disputes over access to some of the world's richest fossil sites are disrupting research on human evolution, according to those working in the field.

Some Ethiopian researchers say new rules have forced them to leave productive sites and stopped them studying their discoveries by denying them laboratory access. And at least one Ethiopian doctoral candidate's degree could be jeopardized by the situation.

Disagreements over site access have been simmering since early 2000. But the dispute heated up considerably last summer when the Ethiopian government introduced new regulations to control physical anthropology projects. The regulations and their impact were the subject of much discussion last month at the annual meetings of the American Association of Physical Anthropologists and the Paleoanthropology Society in Kansas City, Missouri.

The regulations were announced last year by the Ethiopian Authority for Research and Conservation of Cultural Heritage. They set time limits for research, require payments for the use of laboratory facilities — and ask foreign research teams to give up their equipment and vehicles at the end of the project.

But the rules are viewed by some palaeo-anthropologists as biased and damaging to research efforts. And some researchers fear that the rules might not be applied equally to all teams.

Sileshi Semaw, an Ethiopian anthropologist and archaeologist who is now a postdoctoral fellow at Indiana University, believes that the regulations will make it increasingly difficult to do research at the National Museum of Ethiopia and in the field. But he remains optimistic: "Eventually the Ethiopian government will realize what is going on and make a change," he says.

The US National Science Foundation (NSF) is also worried about the regulations, which, officials there say, conflict with NSF policies on the disposal of equipment purchased with the agency's grant money. Last month, Mark Weiss, director of the NSF's physical anthropology programme, met with the Ethiopian ambassador in Washington to discuss the issue. The talks were cordial, says Weiss, with the ambassador saying that concerns would be examined.

Last week, Brook Hailu, deputy Ethiopian ambassador to the United States, said that it is "the general policy of the Ethiopian government that people of high calibre can conduct research". The government will examine the new regulations to see if changes need to be made, said Hailu. "I have no doubt that our government will cancel partially or in total any regulations found to be



Rich pickings: Ethiopian sites are yielding large numbers of hominid fossils.



negative," he added.

There have been plenty of disputes before over research access in Ethiopia, but the present row has arisen over a rewarding site called Galili, about 300 kilometres northeast of Addis Ababa.

Yohannes Haile Selassie, an Ethiopian doctoral candidate at the University of California, Berkeley, spent three years studying this site, finally finding important hominid fossils there in 1998. But in February 2000, an Austrian team headed by Horst Seidler arrived at the site, saying it had government permission to conduct field research there.

Haile Selassie says that the regulations "kicked me out of my site". Since then, he says, he has been unable to conduct analysis on specimens at the National Museum laboratory in Addis Ababa or continue field studies at Galili. His doctoral degree is threatened because he has now exceeded the three-year

limit enforced by the Ethiopian regulations.

When Haile Selassie's mentor, palaeo-anthropologist Tim White, also at Berkeley, stood up for him, the Ethiopian field-research permits for White's Middle Awash research team were suspended in February. White declined to comment.

But Middle Awash team geologist Giday WoldeGabriel, an Ethiopian now at Los Alamos National Laboratory in New Mexico, says that he thinks the regulations favour the Austrian team. "It makes me very angry," he says.

Insisting that he had nothing to do with formulating the Ethiopian regulations, Seidler, of Vienna's Institute of Anthropology, says he has conducted himself appropriately. "Haile Selassie was never displaced from his area," says Seidler, adding that the Ethiopian had not correctly defined his field site. "I have obtained my permit... in a transparent and legal way." ■

PubMed Central offers deal on content

Meredith Wadman, Washington

The website set up by the US National Institutes of Health (NIH) to provide free access to biomedical literature is making moves to quell publishers' unease over the status of online articles.

Organizers of PubMed Central (PMC) now say that participating journals can show free back content on their own websites, rather than displaying full text at the NIH site, as they were previously obliged to do.

PMC's backers say they hope that the policy — announced in an online forum on *Nature's* website last week and published on page 740 of this issue — will encourage many more journals to participate. So far, the NIH site, which was launched in February 2000, has signed up only eight journals. Organizers say that a dozen more are expected to sign up shortly.

"This [new policy] is much less scary for a lot of publishers," says David Lipman, director of the National Center for Biotechnology Information at the NIH. "Some publishers are just concerned about having their content viewed anywhere but their own site." Lipman says the policy change, which was made by the PMC's advisory committee late last month, is "a reasonable compromise".

Journals may now choose to link electronically to PMC, while keeping their content on their own website. In return, they must agree to make original research articles available on their website free of charge within a year of publication, and "preferably" within six months. The change represents a retraction of PMC officials' early insistence that publishers display entire articles at PMC. It comes as several

Funding battle heats up over large array

Jonathan Knight, San Francisco

Astronomers around the world are gearing up to fight to secure funds for a large array telescope in the Andes that would help them study how galaxies are formed.

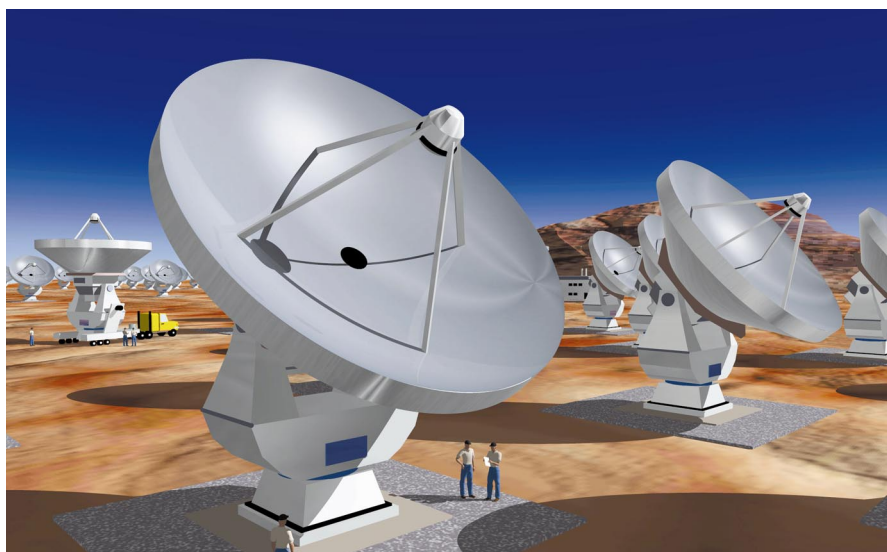
Such an instrument has been a top astronomy priority for years, and on 6 April research-agency representatives from Japan, North America and Europe met in Tokyo but promised only to try to reach agreement on its construction.

Funding for the Atacama Large Millimeter Array (ALMA) is not yet assured — particularly in the United States, which is supposed to pay about one third of its \$660 million construction costs.

The Bush administration's first budget proposal, issued in February, allowed no money for ALMA from the National Science Foundation (NSF), which funds ground-based astronomy in the United States.

The NSF had planned to start construction in 2002, using money it had meant to spend on the smaller Millimeter Array, a US project which was subsumed into ALMA. ALMA also absorbed Europe's Large Southern Array proposal and a Japanese project, the Large Millimeter and Submillimeter Array.

ALMA would be located on the Andean plateau in Chile, where the dry atmosphere will allow short-wavelength signals to reach the telescope. The most distant galaxies can be observed as they were when they first formed because it takes so long for their light to reach Earth. Dust around young stars blocks out most of the visible light in these galaxies, but the millimetre and submillimetre radiation shines through.



Star performer: the Atacama Large Millimeter Array would study the formation of distant galaxies.

"It means that you can make very detailed images of galaxies and stars in formation," says Bob Brown, deputy director of the National Radio Astronomy Observatory, which oversees the US contribution to ALMA.

ALMA's 64 12-metre dishes would simulate a dish of up to 14 kilometres across. Integrating the 64 signals would provide a resolution ten times better than that of the Hubble Space Telescope. As well as resolving very distant galaxies, ALMA should be able to detect planets in our own galaxy.

Funding for ALMA in 2002 is likely to be approved in Europe and Japan, but astronomers face an uphill battle in the United States. A panel of scientists has advised the

NSF that construction should proceed next year.

US researchers are preparing to take their case to Congress, which will consider the Bush proposal and develop a final 2002 budget by October. The Senate has indicated that it would like to give more money to the NSF than Bush proposed (see page 731).

Canada is expected to decide on a US\$30 million contribution later this month, says Ralph Pudritz, an astronomer at McMaster University in Hamilton, Ontario. Pudritz adds that no single country's contribution will be absolutely critical in 2002. "As long as the taps have started to turn on at some significant level, construction can begin," he claims. ■



Lipman: looking for compromise?

journals, including *Science* (see *Nature* 410, 502; 2001), are deciding to post free back content on their own websites, in most cases within a year.

The announcement is the latest development in the debate over how, when and at what price biomedical scientists will be able to access extensive electronic archives of published

articles. Those involved in the discussion, including commercial and academic publishers, rebel scientists and librarians, are to meet next week (16–18 April) at Cold Spring Harbor Laboratory in New York state to discuss electronic publishing issues.

The meeting takes place against the backdrop of a threatened boycott by several thousand scientists, who have pledged to stop buying, publishing in or reviewing for journals that refuse to make their research articles freely available within six months.

Some publishers see the change as an acknowledgement by PMC of a failure to recruit the journals it needs to provide a near-exhaustive archive of biomedical literature. "They've lost that battle, now they're waving a white flag," says one commercial publishing executive, who declined to be identified. "They're saying they'll allow us to not deposit. Well, nobody was going to deposit anyway."

This executive says the PMC's new policy still does not guarantee journals an adequate return. "They are [still] basically trying to pin publishers down to say 'any income we

make we will make in the first six months, at most a year'. That's not a model that most publishers know whether they can accept."

Catherine DeAngelis, editor of the *Journal of the American Medical Association*, calls the new policy "a step in the right direction" because it will allow journals to retain material on their own websites. But she bristles at the demand that journals that link to the site make back content free within a year. "I would love the whole world to get [our content] immediately. But our journal would go under if we didn't have money to run it," she said.

Other publishers say that the change could help the PMC to recruit more journals. Martha Howe, president of the American Society for Microbiology, which publishes ten journals, says that "it may well make participation easier". ■

Biologists bugged by space station cuts and uncertainties

Tony Reichhardt, Washington

The US space agency NASA last week tried to reassure scientists that it intends to furnish the International Space Station with its planned complement of research equipment. But some hardware will arrive years later than planned. And neither NASA nor the White House can yet say how the agency will pay for any science facilities installed after 2003.

Mark Uhlan, director of research integration for the NASA office sponsoring station research, told a task force of the National Academy of Sciences that it was not a question of whether the facilities would be built, "it's a question of when".

His remarks were meant to allay space biologists' fears that some vital hardware — including a centrifuge for investigating the effects of gravity on test animals and plants — might be scrapped (see *Nature* 410, 399; 2001) because of a projected \$4 billion shortfall in the station's budget for the next five years.

The planned US contribution is 27 'racks' of research hardware — multi-purpose facilities each about the size of a fridge-freezer. The first 10 are nearly complete, said Uhlan, and will launch by the end of next year. Partners will contribute their own lab equipment.

When the remaining racks, including facilities for life sciences, fluids and combustion research and materials science, will be installed is less certain. The latest schedule suggests 2004 to 2006. But Uhlan said he could only give a date once NASA had restructured the station's budget in a few months' time.

NASA administrator Dan Goldin told the House Science Committee last week that one option for proceeding after 2004 would be to have European or Japanese partners pay for some lab equipment — perhaps in exchange for more use of the station (the United States has 75% use).

France and Italy are interested in building the crew rescue vehicle and living quarters, both recently cancelled by NASA to save money. Without the rescue vehicle, the station would be limited to three crew members.

Meanwhile, one partner has its own troubles. Japan's space agency NASDA is building the centrifuge for NASA in exchange for a ride on the US space shuttle for Japan's laboratory module. But NASDA managers admit hardware development problems will delay the centrifuge to 2008 at the earliest. ■

Report fudges issue as South Africa fights on against HIV

Michael Cherry, Cape Town

South Africa is set to continue with health policies acknowledging that HIV causes AIDS, after the final report of a government-sponsored panel fudged the question of whether the causal link is proven.

The controversial panel, established last year by President Thabo Mbeki and including many 'dissidents' who dispute the link, produced a report containing several contradictory recommendations, designed to accommodate both sides of the argument.

The report recommends, for example, that testing for HIV infection be suspended "until its relevance is proved especially in an African context, given the evidence of false positive results in an African setting". But it also calls for continued "surveillance of HIV prevalence in antenatal clinics, blood banks and among workers".

Although offering no resolution on the issue, several observers say the report will enable President Mbeki to save face, while allowing the health department to proceed with AIDS policies that acknowledge that HIV causes AIDS.

The panel met twice last summer to consider the causes of the disease. But after its report was released last week, Manto Tshabalala-Msimang, the health minister, conceded that the panel had failed to reach a consensus. She said the government would continue with its existing AIDS programme, while supporting research suggested by the panel into the causes of AIDS (see box).

The director-general in the health department, Ayanda Ntsaluba, declared publicly last October that HIV causes AIDS, and began implementing a pilot programme using antiretroviral drugs to prevent mother-to-child transmission of the disease.

The dissidents appear to be at least partially satisfied by the door being left open for further research.



The way ahead: Tshabalala-Msimang promises that existing anti-HIV measures will continue.

The report's author, Khotso Mokhele, president of the National Research Foundation, recommended last week that another committee be established to consider the merits of these proposals. Tshabalala-Msimang expressed hope that at least some financial support would be forthcoming from outside South Africa.

The long delay in the report's arrival has been the subject of speculation. Initially due at the end of last year, it was handed to Tshabalala-Msimang on 18 January, but she subsequently returned it to the panel's secretariat to deal with what she called "editorial changes of a non-scientific nature".

According to panellist Salim Karim, who is head of the HIV/AIDS unit at the Medical Research Council, "the process of compiling the report was out of the hands of the panellists". ■

Researchers probe the link between virus and AIDS

As part of the South African panel's investigation, a programme of experiments is under way to satisfy any doubts over whether HIV causes AIDS.

In the first of these, blood tests for HIV from five laboratories in South Africa were retested at the US Centers for Disease Control in Atlanta, Georgia. The new tests were 99.9% consistent with the originals — indicating that HIV testing in South Africa is highly reliable.

A second experiment will investigate the claim made by some dissidents that these test results are disrupted by reactions with antibodies raised in response to other diseases, such as malaria and tuberculosis. This work is due to be complete within six to nine months.

A further experiment will use an approach based on the polymerase chain reaction to screen test samples simultaneously for several diseases (or strains thereof) while significantly reducing the risk of contamination. This work is expected to measure the number of multiple infections, and to help identify strains of both tuberculosis and HIV that are resistant to treatment.

M.C.

Bush favours research at Pentagon and NIH

Colin Macilwain, Washington

Military research is up, environmental work is drastically cut, and the National Institutes of Health (NIH) will grow while the rest of civilian science is basically frozen. That is the firmly conservative outline of George W. Bush's first federal budget proposal, details of which were released on 9 April.

The Bush budget for research and development (R&D) reverses the most important trend of the Clinton years, boosting military research by \$2.5 billion — for as yet unspecified purposes — and reducing the civilian share of the \$95 billion R&D spend.

But even as Bush rolled out his numbers on the White House lawn, critics were pointing out that Congress is unlikely to leave many of them intact by the time the 2002 fiscal year begins on 1 October. Just last week, the Senate passed non-binding resolutions that would add more than \$2 billion to Bush's proposals for civilian science.

Announcing plans to increase the NIH's budget by \$2.8 billion to \$23.2 billion next year, Tommy Thompson, the health secretary, pledged that the administration would seek an additional \$4 billion the year after.

Boost for AIDS vaccines

Under the president's plan, the NIH would spend \$12.5 billion on research project grants in 2002, an increase of 13% over 2001. Thompson highlighted the NIH's plans to increase research on AIDS vaccines by more than a quarter to \$357 million, out of \$2.5 billion of research that he said is related to AIDS.

But several scientific groups say that the allocation for the NIH is not enough. Mary Hendrix, president of the Federation of American Societies for Experimental Biology, called it a "great start" and said that the federation would work with Congress to obtain a \$3.4 billion increase by October.

It was a very different story at the National Science Foundation, which funds most non-biomedical research at US universities. Here a basically flat budget of \$4.5 billion left director Rita Colwell with little room for manoeuvre.

As the NSF budget is reshuffled to accommodate some modest new activities, including a \$20 million mathematics initiative, the rest of its research budget will be severely restrained. The amount available for standard, peer-reviewed NSF grants to individual scientists will slip by 0.5%, and the budget for new research equipment falls by a fifth, to \$96 million.

The number of new awards granted during 2002 could actually dip if the agency concludes that NSF grants should be larger and last longer, says Mark Marin of Lewis-Burke Associates, a Washington-based lobby firm whose clients include research universities.



Conservative approach: George W. Bush and his cabinet reveal their proposals for the 2002 budget.

The other highlight in the NSF's budget is a new \$200 million initiative, using money from existing NSF programmes, to tie school education in maths and science more closely with the universities.

Sam Rankin, associate director of the American Mathematical Society, says he is grateful for the \$20 million in the new mathematics research programme — but noted that it falls short of expectations (see *Nature* 407, 931; 2000).

At NASA, a proposed 2% increase will take the budget to \$14.5 billion, although this will not pay for many new initiatives — most of the space agency's activities will be funded at pretty much this year's level. But the agency's five-year budget plan was used by the Bush administration to spell out its space priorities.

Space science fares well in this plan, rising from \$2.45 billion next year to \$3.6 billion in 2006. Most of this increase would go on new missions. NASA Administrator Dan Goldin says the White House cancellation of a near-term Pluto mission — which could yet be reinstated by Congress — "will set us back a little". But he adds that a new advanced propulsion initiative, worth \$310 million over five years, would open new possibilities for exploring the outer planets.

The Department of Energy (DoE), which funds most physics research in the United States, also offered a flat level of funding for next year — to the chagrin of physicists.

"The overall budget is simply not adequate," says Michael Lubell, spokesman for the American Physical Society. With rising power costs and salaries, "a lot of stuff is going to go", he says.

The Bush budget proposal allows no increase at all for nuclear physics and just over 1% increase for high-energy physics

and basic energy sciences.

Lobbyists complain that the budget does not match Bush's statement — speaking to newspaper editors on 9 April — that "basic research gets big increases" in his budget. "They ought to have the guts to say what they're doing," says Lubell.

Energetic cuts

But other sciences fare even worse at the DoE, with the biological and environmental research division losing 8% of its 2001 budget. Energy secretary Spencer Abraham also cut funding for renewable energy technologies by 36%, observing that they were "not in focus with where industry was headed".

The department's nuclear weapons research programmes would grow by about 5%, to \$5.3 billion, with Abraham pledging to resume production of plutonium triggers, which stopped in 1989. Programmes aimed at restraining nuclear proliferation, including aid to Russian weapons labs, are cut by 12%.

The rest of the budget contains aggressive cuts in environmental science. Research at the Environmental Protection Agency, for example, is reduced by more than a fifth, to \$707 million. The agency's research on clean air and clean water is cut back by a third.

The National Oceanic and Atmospheric Administration — an important player in climate change and ocean research — loses 10% of the \$365 million that its main science office has to spend this year.

And the United States Geological Survey loses about 8% of its 2001 budget. Its biology division falls from \$164 million to \$152 million and its water resources division plunges by almost a quarter, from \$202 million to \$159 million. ■

Additional reporting by Matthew Davis, Corie Lok, Tony Reichhardt and Paul Smaglik.

Professors seek funds to fight AIDS in developing countries

Washington A group of professors at Harvard University have called on developed countries to provide around \$6 billion a year to pay for AIDS treatment and prevention programmes in developing nations.

In a signed statement released last week, 128 academics called for wealthy countries to set up a global trust fund to ensure that AIDS victims in poorer countries have access to the drugs they need. The professors claim that providing antiretroviral therapy for three million AIDS sufferers would cost \$3.3 billion annually. A further \$3 billion is needed for other drugs and prevention programmes.

On 5 April, the US Senate approved a non-binding resolution from senator Bill Frist (Republican, Tennessee) that would double current US spending on international HIV/AIDS programmes to more than \$1 billion over the next two years.

European labs join forces to beef up food safety

Paris Four of Europe's principal food safety research institutes are pooling resources to form Safety for Food in Europe (SAFE). The

consortium aims to capitalize on funding that will be available next year from the planned new European Food Authority (EFA).

The EFA will be the new centrepiece of the European Commission's work on food policy and will oversee all aspects of food safety "from farm to fork". SAFE hopes to use EFA funding to conduct joint research programmes and to provide information for consumers.

The founding members of SAFE are the Organization for Applied Scientific Research and the University of Wageningen, both in the Netherlands, the French agronomic research agency INRA, and the Institute of Food Research, based in Norwich, UK.

Conservative to head German research body

Munich The Gottfried Wilhelm Leibniz Association of German Research Centres (WGL) has appointed a controversial new chairman.

Hans-Olaf Henkel, a prominent figure in German political life and former head of the Association of German Industry, is regarded as a firm conservative. The WGL — whose members include the cream of the former East German research institutes, such as the Potsdam Institute for Climate Impact Research and the Institute of Molecular



Henkel: keen to boost competition.

Biology in Jena — has been criticized for its reluctance to close poorly performing facilities. By appointing Henkel, the WGL hopes to demonstrate that it is serious about raising standards.

Henkel is the first non-scientist to chair a large German research organization. He will

take up office in July, but has already announced plans to increase competition among WGL members by setting up a joint strategy fund.

Czech science academy names female president

Munich Helena Illnerová of the Institute of Physiology in Prague has become the first woman to be elected president of the Czech Academy of Sciences. The 63-year-old neuroscientist is best known for her research on processes in the brain that control daily cycles of sleeping and waking.

The appointment of Illnerová follows a turbulent time for the Czech academy. Her predecessor, Rudolf Zahradník, closed 20 institutes and halved staff numbers to around

6,000. Illnerová looks set to make further changes, and plans to carry out regular external reviews of the performance of the remaining 62 institutes.

But Illnerová says that her main goal will be to increase support for the next generation of Czech scientists. She hopes to introduce a programme to provide extra money for the country's most talented young researchers by next year.

Animal-rights success scares UK companies

London Five major pharmaceutical companies have written to Jack Straw, the British home secretary, to express concern over the actions of animal-rights activists.

The companies themselves have not gone public over the letters, but their existence was revealed last week by Trevor Jones, director-general of the Association of the British Pharmaceutical Industry (ABPI). Jones claims that ABPI members are worried about the success of an animal-rights campaign against UK drug-testing laboratory Huntingdon Life Sciences. Activists have targeted financial and pharmaceutical companies that work with Huntingdon, forcing them to sever links with the laboratory.

The ABPI said it welcomed recent plans by

Straw for new laws designed to protect company directors and employees from intimidation by activists, but stressed that the government had to ensure the laws worked in practice.

Testing times as insurers are slammed over genetics

London A UK parliamentary committee has called for a two-year moratorium on the use of genetic testing by insurance companies. In a damning report issued last week, the House of Commons' Science and Technology Committee concluded that insurers were "more interested in establishing their future rights to use genetic test results in assessing premiums, than in whether or not they are reliable or relevant".

The committee found that some insurance companies were using unapproved tests to set levels for premiums. They have been testing for conditions such as hereditary breast cancer and ovarian cancer, and gave the results a "predictive significance that cannot at present be justified".

The UK government's regulatory body currently allows life insurers to use the results of genetic tests only for Huntington's disease in setting premiums. It is considering also permitting tests for early onset Alzheimer's disease to be used.

Successful launch for NASA Mars mission

Washington NASA set about getting its Mars exploration programme back on track with the launch of its first Mars mission since two spacecraft went missing on arrival at the Red Planet in 1999.

The 2001 Mars Odyssey took off successfully from its Florida launch pad on 7 April, and is scheduled to arrive in Mars orbit in October. During its 30-month mission, the \$300 million spacecraft will search for evidence of water and ice beneath the surface. The results are certain to influence the debate about whether there was life on Mars. The spacecraft will also analyse mineral levels on the planet and study levels of ultraviolet radiation reaching the surface.

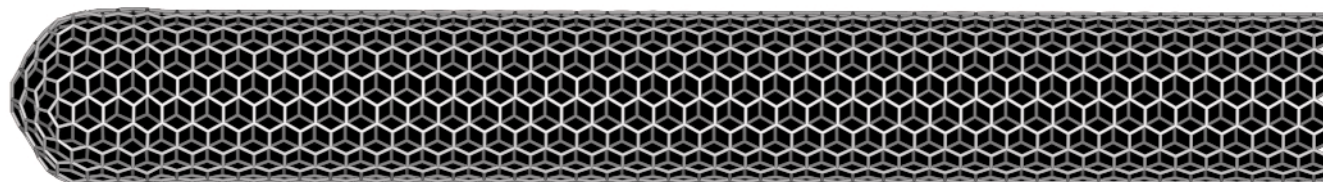
NASA has not included a lander this time, but the agency's next Mars mission, in 2003, will carry a pair of robotic rovers that will explore the surface of the planet.



Blast off: the 2001 Mars Odyssey.

Is it all just a pipe dream?

Researchers trying to turn nanotubes into storage systems for hydrogen fuel are finding that corporate funding and academic openness can be hard to combine. Catherine Zandonella delves into a carbon controversy.



The fuzzy black soot at the bottom of the flasks in Michael Heben's lab looks ordinary enough. But for those funding his work at the National Renewable Energy Laboratory in Golden, Colorado, it might someday be worth a fortune. If Heben is right, this grimy substance might be able to store fuel for a new generation of hydrogen-powered cars. But with money and patents at stake, some researchers feel that academic openness is playing second fiddle to corporate interests.

The soot contains carbon nanotubes, tiny hollow cylinders around 1.5 nanometres in diameter. Their walls are made up of a single layer of carbon atoms, and can be thought of as a sheet of atoms arranged in a hexagonal pattern and then rolled up to form a tube — much like a miniature roll of chicken wire (see above). Based on the nanotubes first synthesized in 1991 by Sumio Iijima at the NEC Corporation in Japan¹, these ultra-strong, lightweight tubes have potential uses ranging from wiring in integrated circuits to components in nanoscale motors. And for researchers in the clean-energy business, nanotubes carry the tantalizing potential of being a new storage system for hydrogen fuel.

Billed as the energy source of the future, hydrogen is plentiful and clean. Fuel cells — batteries that combine oxygen and hydrogen to produce electricity — could one day power virtually all cars and buses, pumping out nothing more harmful than warm water. But for hydrogen to be a practical fuel for cars of the future, researchers must first develop a lighter storage system to replace the bulky tanks of liquefied or condensed hydrogen found in many of today's electric vehicles.

Carbon nanotubes have the right qualifications for the job. They can be arranged in groups to form rope-like bundles. The diameter of the individual tubes is typically two or three times that of a hydrogen atom, and researchers believe that hydrogen could be tucked both inside and between the tubes. The amount of hydrogen that nanotubes can

adsorb in this way depends on pressure and temperature, so changing these conditions should allow hydrogen to be pumped in and out of the system.

Test tubes

Despite this exciting potential, years of work on nanotubes have produced mixed results. Progress has been made, but funding arrangements are muddying the waters. Some researchers using corporate money are prohibited from fully disclosing their methods until patents are filed, generating doubts about their findings. The simmering tensions came to a head last November at a Materials Research Society (MRS) meeting in Boston. Acrimonious debate erupted after some researchers publicly criticized new results.

The history of hydrogen storage in carbon materials is littered with promising findings that seemed to collapse under closer scrutiny. One frequently cited example is the work of Nelly Rodriguez, a researcher at Northeast-

ern University near Boston. In 1998, Rodriguez announced that she could get graphite nanofibres — thin wires of carbon — to adsorb their own weight in hydrogen². Citing patent concerns, she refused to reveal the details of her technique, generating widespread scepticism about her results.

So far, no one has managed to reproduce Rodriguez's results³. She, in turn, has stopped presenting her findings at meetings because of the criticism she received. Rodriguez is now leaving Northeastern to concentrate on her company, Catalytic Materials, based in Mansfield, Massachusetts, which has the backing of several unnamed corporations. "The project is going extremely well," she says, "but we cannot publish our work because we have contracts that are proprietary. This is one of the reasons we are leaving academia."

Another example is work conducted by Jianyi Lin at the National University of Singapore. In a 1999 paper, Lin and colleagues reported high storage levels in lithium-doped



Up in smoke: pollution from today's cars could be eliminated if hydrogen fuel can be stored effectively.

multiwalled nanotubes⁴ — tubes that are nested one inside the other like Russian dolls. Lin claimed that 20% of the combined weight of nanotubes and adsorbed hydrogen was attributable to the hydrogen — described as 20% by weight. Since then, other researchers have shown that Lin was actually measuring storage of water rather than hydrogen⁵.

But despite these incidents, new results have continued to attract the interest of energy and car companies. A more practical hydrogen storage system would bring huge financial rewards, so corporations are willing to forgive a few false alarms.

Carbon conundrums

To foster research in this area, the US Department of Energy (DoE) has set target storage levels for vehicle fuel systems. It says that for a practical system, a tank should hold 62 kg of hydrogen per cubic metre, or 6.5% by weight. The system needs to adsorb and release hydrogen quickly while working at temperatures ranging from 25 to 100 °C, and preferably at pressures of no more than 200 times normal atmospheric pressure.

Since the guidelines were issued, Heben's team has used a new technique to produce purer nanotubes⁶, eliminating virtually all the carbon not in the form of tubes from the samples. At the International Winterschool on Electronic Properties of Novel Materials, held in March last year in Kirchberg, Austria, he presented spectacular results, beating the DoE goals with storage of up to 7% hydrogen by weight at room temperature and ambient pressure in tubes he says are up to 98% pure.

But at November's MRS meeting, Michael Hirscher of the Max Planck Institute for Metals Research in Stuttgart announced that he had repeated Heben's experiments using similar nanotubes and obtained very different results. Not only did Hirscher dispute Heben's figure of 7% storage — his nanotubes could only store 1.5% hydrogen by weight — but he found traces of titanium contamination within the tubes. Titanium adsorbs hydrogen, and Hirscher claimed that the metal was responsible for all of the adsorption. The titanium came from a probe used to open the ends of the tubes in an effort to increase their storage capacity. When Hirscher used a stainless-steel probe, the tubes' storage capacity plummeted. Hirscher backed up his claims this February, when he published an invited paper⁷ in the journal *Applied Physics A*.

Heben was astounded at what he saw as a betrayal of a working relationship with Hirscher. The two had been in contact for some time, and Heben had supplied Hirscher with several key details of the experiment. Heben says that he is well aware of the titanium. Rather than being a contaminant, Heben sees it as vital to the higher storage density — but not in the way that Hirscher suggests. Heben agrees that the titanium may adsorb small quantities of hydrogen. But more



Tanked up: electric buses promote clean air, but require bulky containers to store hydrogen fuel.

importantly, he believes that its presence somehow stimulates the carbon nanotubes to take up more hydrogen — although he admits that he cannot prove this yet. Heben now has a patent for using titanium in carbon nanotubes. As for Hirscher's recent paper, Heben claims that it is flawed and points out that it was not sent out for external peer review.

Hirscher's results, says Heben, may simply have been caused by using inferior nanotubes. "Some people think nanotubes are all the same," says Heben, "but our purity is much higher." The matter might be settled if Hirscher used exactly the same tubes as Heben, but Heben is reluctant to release his nanotubes until he obtains consistent results for hydrogen storage. At the moment, the value he gets varies depending on how it is measured — and when he uses the method favoured by most other researchers, he gets only up to 4% by weight. But even if Heben does resolve this issue, he says that the bad feeling between the two parties is such that he may never be willing to supply his materials to Hirscher.

Commercial brakes

Sponsorship from industry has again complicated matters. Heben is involved in a public-private partnership designed to speed the transfer of DoE-sponsored research to commercialization. Under the terms of his agreement with Honda R&D Americas, Heben must secure intellectual property rights

before releasing full details of his work. Some researchers suspect that this has prevented Heben from disclosing all the details of his technique for making pure nanotubes.

Heben admits that the contract "breeds mistrust", but denies that he is hiding anything. He says that by securing a patent before he announced his results, he was able to be completely open about his work. As someone who worked on experiments that failed to replicate Stanley Pons' and Martin Fleischmann's 1989 claim to have achieved cold fusion, Heben knows all about controversy. "We just want to find out what works," he says, "and if we find out it is wrong, so be it."

As Heben's salary depends heavily on funding from the DoE and his corporate partners, this attitude seems commendable. Without promising results, his funding could be decreased or withdrawn.

Doubts will remain until Heben allows other labs to test his nanotubes. But with no one else close to meeting the DoE targets, some scientists are beginning to lose faith. Many European researchers seem convinced by the finality of Hirscher's work, although some Japanese and American groups remain open-minded. "I think it is too early to conclude this story," says NEC's Iijima.

Heben will give hydrogen vehicles a big boost if he can prove his critics wrong. And if he fails, at least everyone involved will have learnt a little more about the tricky business of mixing corporate money with academic-style research.

Catherine Zandonella is a science writer in San Francisco.

A more practical hydrogen storage system would bring huge financial rewards.

1. Iijima, S. *Nature* **354**, 56–58 (1991).
2. Chambers, A., Park, C., Baker, R. T. K. & Rodriguez, N. M. *J. Phys. Chem. B* **102**, 4253–4256 (1998).
3. Ahn, C. C. et al. *Appl. Phys. Lett.* **73**, 3378–3380 (1998).
4. Chen, P., Wu X., Lin, J. & Tan, K. L. *Science* **285**, 91–93 (1999).
5. Yang, R. T. *Carbon* **38**, 623–626 (2000).
6. Dillon, A. C. et al. *Adv. Mater.* **11**, 1354–1358 (1999).
7. Hirscher, M. et al. *Appl. Phys. A* **72**, 129–132 (2001).

Making crops cry for help

Plants attacked by hungry herbivores can release chemicals that attract their assailants' predators. Could these responses be exploited to develop environmentally friendly pest-control strategies? John Whitfield investigates.

Plants cannot run from trouble, but neither do they lie down and surrender. As well as producing a variety of noxious chemicals to deter herbivores, they can enlist help from higher up the food chain, releasing volatile chemicals that attract predators to eat the creatures that are eating them.

Some experts in crop protection believe it may be possible to tap into these lines of communication to provide non-toxic forms of pest control. But a great deal of research into the chemistry, genetics and ecology of these botanical strategies for help will be needed to bring such strategies to the market. "It's definitely in the long-term category, but a very far-thinking company could put some effort into this," says Jonathan Gershenzon, who works on the genetics of plant defences at the Max Planck Institute for Chemical Ecology in Jena, Germany.

When plants are attacked by pests, they release volatile organic chemicals that attract predators and parasitoids — insects that lay their eggs inside other insect larvae, which are then devoured from within. The production of these volatiles was first noticed more than a dozen years ago¹. Subsequent studies have revealed that predators and parasitoids learn to associate the volatile chemicals with the presence of their prey and hosts, and that plants usually produce the signals as a



Double act: maize can be protected from pests by planting 'intercrops' that attract predators.

response to chemicals in the pest species' saliva, rather than to mechanical damage^{2,3}. Their production can also be induced by a plant hormone called jasmonate⁴.

The simplest approach would be to spray the volatile chemicals directly onto crops, or to induce their production using jasmonate. And in 1999, when Jennifer Thaler of the University of California, Davis, sprayed Californian tomato fields with jasmonate, she found that parasitism of *Spodoptera exigua* caterpillars by *Hyposoter exiguae* wasps increased twofold⁵.

Crying wolf

Despite this success, many researchers remain sceptical of the value of such simple spraying regimes. "You don't want to make the mistake of attracting a parasitoid without a host being present," says Consuelo De Moraes of the US Department of Agriculture's Center for Medical, Agricultural, and Veterinary Entomology (CMAVE) in Gainesville, Florida. Farmers do not like to wait for an infestation to become established before applying pesticide sprays. But crying wolf by spraying preventatively would lead to the learned association between the plant volatiles and the availability of host species being broken — and the chemicals would soon lose their attractant power.



Fatal attraction: spraying crops with the plant hormone jasmonate attracts wasps that parasitize *Spodoptera exigua* caterpillars.

Instead, many experts believe a better option will be to produce plant varieties that respond more strongly to attack by pests. "We are not there yet, but we shall be heading in that direction in the next few years," says James Tumlinson, a chemist at the CMAVE. Already, his team has identified a wild strain of cotton that produces up to ten times more parasitoid-attractants than domesticated varieties⁶. Over the past century, says Tumlinson, ignorance of this subtle chemistry seems to have resulted in plant breeders presiding over a decline in cotton's ability to call for help.

Such traits could be introduced into crops by conventional plant breeding. And in the Netherlands, that approach is being considered to protect the country's flower-growing industry. Marcel Dicke of Wageningen University is working with plant breeding companies and the Dutch Technology Foundation to develop varieties of the daisy *Gerbera* that are more resistant to herbivorous mites and more attractive to predatory mites. Dicke aims to gain a better molecular understanding of the volatiles involved, and how they are produced. Some of the enzymes involved in volatile manufacture are known^{7,8}, and it ought to be possible to speed up the process of selecting plants to be used for breeding by screening cultivars for enzymatic activity. This would bypass laborious tests involving exposure to pests.

Dicke also intends to use DNA microarrays to determine patterns of gene expression in plants responding to herbivore attack. His Wageningen team is just one of several groups turning their attention to the model plant *Arabidopsis thaliana*, in the hope that it will yield insights that can be applied to crops.

Volatile response

Such studies might suggest ways of genetically engineering plants to boost their production of volatiles in response to herbivore attack. "In principle, it shouldn't be difficult," says Gershenzon, who plans to study maize plants in which genes suspected to be involved have either been knocked out or engineered to produce larger quantities of their products. But although genetic engineering would probably be more efficient than conventional breeding, those gains would have to be balanced against public suspicion of plant genetic modification.

There is also much work to be done to understand how parasitoids and predators interpret the complex cocktails of volatiles produced by plants attacked by herbivores. Varieties of maize, for example, release blends with a dozen or more major and minor chemical components⁹. Which of these are key to the parasitoid response is unclear. Also unknown is how parasitoids learn to associate plant volatiles with their hosts. "It may not matter what the plant releases," suggests Gershenzon, "as long as it's distinctive, and released quickly in large amounts."

So far, industry has not come forward with large sums of money to invest in the necessary research. "If I were a plant breeder, I'd be saying 'show me that it works in the field,'" says Ian Baldwin, who also works at the Max Planck Institute in Jena.

But Baldwin and his colleague André Kessler are working to address this scepticism. Last month, they reported on studies of wild tobacco, *Nicotiana attenuata*, in Utah¹⁰. This plant is attacked by caterpillars of the moth *Manduca quinquemaculata*, the eggs of

If you're a farmer trying to use natural enemies, it's important to think about where they're coming from.

which are eaten by the bug *Geocoris pallens*. By studying the volatiles released by the tobacco plants and applying them individually and in combination to the stems of plants growing in the field, Kessler and Baldwin identified three compounds that increased predation of *M. quinquemaculata* eggs by *G. pallens*. A blend of the three also reduced egg-laying by the moth. From these experiments, the researchers estimated that, by releasing volatiles, tobacco plants could reduce the number of herbivores attacking them by more than 90%.

One especially hopeful sign is that *G. pallens* is not a fussy eater, and will prey on a wide range of herbivorous species. To Baldwin, this suggests that we may not need to understand the fine ecological details of every plant-herbivore-predator interaction to make use of plant volatiles in crop protection.

But Baldwin's studies fall short of showing that crops that produce more volatiles will attract enough extra predators to give effective and economically viable protection from pests. That would depend on there being a nearby reservoir of predators and parasitoids to rush to the plants' defence.

This is a real concern in intensive agricultural systems, where crops are grown in monocultures in enormous fields that may lack the necessary biodiversity. "If you're a farmer trying to use natural enemies, it's important to think about where they're coming from," says Thaler, who is now at the University of Toronto. "Stealing them from your neighbour's field isn't going to work." Indeed, some researchers suggest that the success of approaches to pest control that use plant-predator communication may depend upon the existence of 'set-aside' areas, hedgerows or unplanted field edges, to harbour populations of predators and parasitoids.

Mix and match

In the developing world, where agriculture does not function so much like an industrial process, this thinking can be extended to develop mixtures of crops that can be grown together to encourage predators and parasitoids. In East Africa, the staple crops of maize and sorghum are under constant attack from stem-borer caterpillars, slashing yields by up to 80%. African farmers have



Pest busters: herbivores such as this white spider mite can be curbed by their predatory relatives.

SPL

little or no access to chemical pesticides, but it is relatively easy for them to adapt their practices to plant and harvest mixtures of species.

With this in mind, researchers from the International Centre of Insect Physiology and Ecology in Nairobi, Kenya, and Britain's Institute of Arable Crops Research (IACR) at Rothamsted in Hertfordshire, have designed a mix of plants dubbed 'push-pull'. Around fields of maize and sorghum they plant species that the stem-borers like to eat, pulling the pests from the crop. Among the crops, they grow species that repel stem-borers and attract parasitoids, including one species, *Melinis minutiflora* or molasses grass, that releases parasitoid-attractants even when untouched by pests. In such fields, the number of plants infested with stem-borers drops by more than 80%, and the number of parasitized larvae rises almost fourfold¹¹.

Following these successful Kenyan trials, says John Pickett of the IACR, push-pull planting is gaining popularity. The government in Uganda plans to promote it, Malawi is interested, and farmers in Ethiopia have taken it up of their own accord. There is also interest from China, India and countries in South America.

Such strategies may not yet be applicable in the developed world, but they show that plant-predator communication can be exploited in pest control. "There are a lot of unanswered questions, but there's a lot of potential," says Gershenzon. ■

John Whitfield works in Nature's science writing team.

1. Dicke, M. & Sabelis, M. W. *Neth. J. Zool.* **38**, 148–165 (1988).
2. Turlings, T. C. J. *et al. Proc. Natl Acad. Sci. USA* **92**, 4169–4174 (1995).
3. Takabayashi, J. & Dicke, M. *Trends Plant Sci.* **1**, 109–113 (1996).
4. Thaler, J. S., Stout, M. J., Karban, R. & Duffey, S. S. *J. Chem. Ecol.* **22**, 1767–1781 (1996).
5. Thaler, J. S. *Nature* **399**, 686–688 (1999).
6. Loughrin, J. H. *et al. J. Chem. Ecol.* **21**, 1217–1226 (1995).
7. Bouwmeester, H. J., Verstappen, F., Posthumus, M. A. & Dicke, M. *Plant Physiol.* **121**, 173–180 (1999).
8. Degenhardt, J. & Gershenzon, J. *Planta* **210**, 815–822 (2000).
9. Turlings, T. C. J., Lengwiler, U. B., Bernasconi, M. L. & Wechsler, D. *Planta* **207**, 146–152 (1998).
10. Kessler, A. & Baldwin, I. T. *Science* **291**, 2141–2144 (2001).
11. Khan, Z. R. *et al. Nature* **388**, 631–632 (1997).

Silicon philanthropists follow a great tradition

Sir—What is being called ‘venture philanthropy’ in your News feature ‘Biomedical philanthropy, Silicon Valley style’ (*Nature* **410**, 140–143; 2001) is hardly a new concept — it is an example of what has traditionally characterized ‘thoughtful philanthropy’. In fact, it is very like the ‘old’ strategic philanthropy envisioned and practised by many private foundations established at the beginning of the twentieth century.

Robert Kohler’s book *Partners in Science* (Univ. Chicago Press, 1991) provides many examples of the essential role private funders played in the development of modern molecular biology, genetics, public health and other biomedical research fields. The Silicon Valley entrepreneurs interviewed in your feature are following a tradition long established by private funders who viewed their money as providing venture capital for the common good and who long held the view that philanthropy should invest in the acquisition of new knowledge and in its responsible application. Further, the staff of the new foundations are following in the footsteps of a profession first defined by the Rockefeller Foundation ‘circuit riders’ — programme officers knowledgeable in their fields, actively seeking out promising research projects needing support.

I do not intend to detract from the laudable efforts of these new philanthropists to pursue thoughtful giving that takes risks and invests in new ideas with a minimum of hassle and red tape. But I do not see why they need to be flattered into thinking they are inventing something new — or that, as some of them seem to believe, they have a new model that needs to be emulated by all funders. Just as there is a certain dishonesty in the communication between researchers and government funders (the ‘proposal’ describing work already completed), there is a growing risk of dishonest dialogue between scientists and some of the philanthropists identified by the anecdotes Trisha Gura relates (‘No one but you has the keen insight to recognize my brilliant idea’).

The number of projects and researchers supported by private philanthropy will remain small compared with those receiving government support, and private funding relies on its partnership with public dollars. Peer review may not be perfect, but neither is investing in whoever grabs someone’s attention or ear.

The true richness of private philanthropy is found in its diversity of

approaches and its distributed decision-making processes that allow many different points of view — and many different grant-making approaches — to flourish.

Susan M. Fitzpatrick

James S. McDonnell Foundation, 1034 South Brentwood Boulevard, Suite 1850, St Louis, Missouri 63110, USA

When DNA research menaces diversity

Sir—The *Nature Science Update* “Synthetic viruses just around the corner”¹ reported a discussion about scientists’ imminent ability to synthesize new viruses for producing better vaccines or for devising deadlier biological weapons. I would like to give an example where genetic engineering of viruses can undermine careful management of natural resources.

Two new rabbit virus strains are being developed for opposing reasons. On the one hand, a group in Australia is modifying rabbit myxoma virus to transmit reproductive sterility and so reduce numbers in a region where rabbits constitute a pest². On the other, a European group is modifying myxoma virus to express rabbit haemorrhagic disease virus capsid protein³ to protect rabbits against both diseases and encourage the recovery of wild populations within the species’ original distribution area.

Each goal is logical within its regional context. However, the history of rabbit viruses shows that outcomes can conflict with initial goals. First recorded in Montevideo in 1896, myxoma virus was successfully released in Australia in 1950 to control rabbit populations, but an illegal release in 1952 in France led to the virus spreading throughout Europe. Similarly, rabbit haemorrhagic disease, which spread naturally from Chinese rabbitries throughout Europe in 1987, spread in Australia after an accidental release from a trial island and in New Zealand after an illegal release in 1997.

The new modified viruses could spread worldwide as easily as myxomatosis and rabbit haemorrhagic disease initially spread. The establishment of modified myxoma virus into inappropriate regions could have disastrous effects on biodiversity. The preservation of Australasian ecosystems, as well as the conservation of endangered predators in Europe, depends on the same species: the wild rabbit, the target of both modified viruses.

It is therefore essential that modified viruses are very carefully and appropriately used. Biotechnology policies on the release of modified organisms rely on national authorities. But rabbits are distributed

throughout the world, so it is essential to guide the development, release and regulation of rabbit virus biotechnology, and to enforce international controls to prevent accidental spread of genetically modified viruses.

Elena Angulo

Instituto de Investigación en Recursos Cinegéticos, Apdo. 535, E-13080, Ciudad Real, Spain, and Estación Biológica de Doñana, Apdo. 1056, E-41080 Sevilla, Spain

1. Whitfield, J. *Nature Science Update*. <http://www.nature.com/nsu/010222/010222-17.html> (accessed 21 February 2001).

2. Kerr, P. J. *et al. Biol. Reprod.* **61**, 606–613 (1999).

3. Bárcena, J. *et al. J. Virol.* **74**, 1114–1123 (2000).

Gulf syndrome research has passed peer review

Sir—Your News report (*Nature* **410**, 135; 2001) mentioned the negative comments made by a former US Department of Defense (DoD) employee, Bernard Rostker, and by Philip Landrigan, a Clinton committee appointee, about our published research on Gulf War syndrome. Because your story stated that our work was supported by a private foundation and non-peer-reviewed grant funds, and involved small patient samples, readers might be left with the impression that our findings are invalid.

This is not true. Over the past four years we have published 12 papers in prominent scientific journals establishing that there is a new syndrome with three variants in Gulf War veterans of a naval reserve battalion (see, for example, ref. 1). Our study of 63 cases and 186 controls identified strong associations with risk factors for exposure to sarin nerve gas and related chemicals. We have also identified affected brain regions and a genetic predisposition using a variety of techniques. Our design and sample sizes are equivalent to those used by the Centers for Disease Control and Prevention in such classic epidemic investigations as toxic shock syndrome, Four-Corners hantavirus pneumonia and AIDS.

The funding proposals for all our studies were rejected by the DoD’s peer-review system but were funded by a private foundation or after appeal to higher government levels. Our results later passed rigorous peer review before publication in respected scientific journals. A consistent publishing record is a better indicator of scientific merit than the sources of funding, particularly in the politically charged environment of Gulf War research.

Robert W. Haley

University of Texas Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, Texas 75390-8874, USA

1. Haley, R. W. *et al. Arch. Neurol.* **57**, 1280–1285 (2000).

PubMed Central decentralized

Sir — In 1999 the US National Institutes of Health (NIH) suggested that a freely accessible public archive for the scientific literature would greatly benefit the scientific community. To date, more than 20 journals have contributed material to PubMed Central, the electronic archive born from that NIH proposal. To encourage wider participation, PubMed Central (<http://www.pubmedcentral.nih.gov/>) is now offering publishers the option of depositing material for archival purposes without making the full text viewable. Searches for this material at PubMed Central would lead users to the full-text articles at the publisher's site.

Life-science publishing, like any other consumer industry, has had to respond to the technological change brought about by the Internet. A large, growing proportion of science journals now publish online versions; online-only journals are sprouting up; and several journals now take online submissions and use online peer review.

Two years ago, Harold Varmus, then director of the NIH, announced the E-biomed initiative to ensure a robust electronic archive of life-science research articles, freely accessible to everyone. After much discussion and the incorporation of the ideas of many people with diverse interests, E-biomed became PubMed Central, the free life-sciences journal archive. It is managed by the National Center for Biotechnology Information (NCBI), a part of the US National Library of Medicine (NLM), which has significant experience in the creation of online archives, exemplified by PubMed (MEDLINE) for biomedical abstracts and GenBank, for nucleotide sequences.

As things stand, participating journals submit their content to PubMed Central either at the time of publication or after a lag of up to a year or more after publication. PubMed Central archives the articles and makes them universally available to readers. Until now, participating publishers had to allow their full text to be displayed free at the PubMed Central site where it could be searched or manipulated, for example to create links to GenBank. Our new option means that a publisher can now stipulate that full text is seen only at its own site, with the publisher's own restrictions on access. PubMed Central's sole condition is that this content is made available free at the publisher's site within one year of publication (preferably within six

months). If the publisher fails to comply, the NLM will have the right to make the material freely viewable in PubMed Central a year after publication.

Publishers choosing the new option will submit their full text to PubMed Central as they do now, in SGML or XML (mark-up languages) files conforming to a document-type definition (DTD — a mark-up template) for journal articles, and as high-resolution image files. Viewable and non-viewable submitted articles must meet a PubMed Central standard to ensure the integrity of the archive (see 'rules of the game' in our contribution "PubMed Central decides to decentralize" at <http://www.nature.com/nature/debates/e-access/>).

Over the past few months the PubMed Central archive has developed a new software architecture, built around the concept of a common template and precise specification for data tagging. This new DTD, based on the latest XML standards, creates a more detailed and sustainable archival copy of an article (see figure) than HTML, which currently serves as the 'standard' electronic record for most online journals.

The new standard DTD means that every article in the archive has its parts (authors, affiliations, text, references, and so on) tagged in exactly the same way, regardless of its source format. (The first 25 journals using PubMed Central have 10 different DTDs.) PubMed Central's normalized tagging, which has no effect on the article's content, greatly simplifies all further use of the archive. It allows, for example, searches for reagents mentioned only in the methods section of an article or searches of just the figure legends. It simplifies the provision of any feature that depends on knowing the context of a string of terms in an article. Further integration of the content of PubMed Central journals with other NCBI resources — such as genomes, macromolecular structures and online textbooks — is also facilitated.

As with any new kid on the block, we may not yet have as broad a set of capabilities as the more established players. Much of our recent emphasis has been on the unglamorous but essential work of creating a stable, robust archive architecture. By starting later, we have greater flexibility to adopt newer technologies and therefore to introduce new capabilities.

Many myths have arisen round PubMed Central. It does not publish preprints and other unrefereed material; participation does not involve any

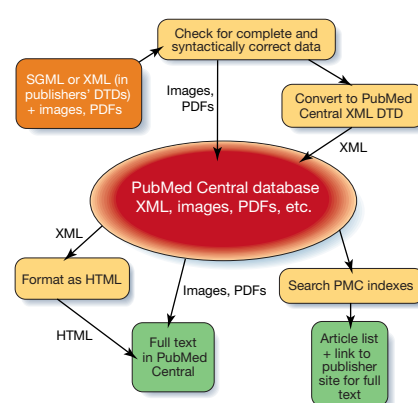


Figure 1 PubMed Central data flow.

'exclusive use' agreement; and PubMed Central does not charge for archiving or related services (see 'Dispelling the myths' in "PubMed Central decides to decentralize" at <http://www.nature.com/nature/debates/e-access/> for more details). The NLM has been collecting and preserving the medical literature for more than a century; the extension to stewardship of the electronic literature is a natural step.

But what does preservation have to do with free access? The only way to ensure the permanence of an electronic archive is to use it continuously, and there is no better way to do that than making it freely available to everyone. PubMed Central was built on the twin standards of a permanent archive and free access. We are now stretching the latter principle to direct users to a publisher's site for full text. We hope that many more publishers will thereby be encouraged to contribute to the archive so it can realize its full potential — in ways still to be discovered.

Edwin Sequeira, Johanna McEntyre, David Lipman

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

● This article is a slightly edited version of a contribution to *Nature's* current web debate on electronic publishing initiatives in science (see *Nature* 410, 613; 2001). Readers wishing to participate in the debate by replying to this or any of the other contributions are invited to view (<http://www.nature.com/nature/debates/e-access/>). Contributions can also be submitted to Correspondence via corres@nature.com. In either case, publication will be offered according to the criteria described on page 613 of last week's issue — Correspondence Editor, *Nature*.

Science for the have-nots

Developed and developing nations can build better partnerships.

Ahmed H. Zewail

Only a fifth of the population enjoys the benefit of life in the 'developed world', and the gap between the haves and have-nots continues to increase, threatening stability. According to the World Bank, of the 6 billion people on Earth, 4.8 billion live in developing countries, 3 billion live on less than US\$2 a day, and 1.2 billion live on less than \$1 a day, which defines the absolute poverty standard; 1.5 billion people do not have access to clean water.

Although globalization, in principle, aspires to help nations prosper and advance, it is better tailored to the fraction of the world's population able to exploit natural resources and markets. The per capita gross domestic product — the total unduplicated output of economic goods and services produced within a country as measured in monetary terms — has reached \$35,000 in some Western countries compared with about \$1,000 in many developing countries, and significantly less in underdeveloped populations. For example, according to United Nations statistics: Angola \$528; China \$777; North Korea \$430; South Korea \$6,956; United States \$31,059; and Yemen \$354. This situation will, if left unchecked, exacerbate global instability.

Progress barriers

Developing nations encounter four barriers to achieving developed-world status. High rates of illiteracy in many developing countries reflect the failure of educational systems, and are linked to the alarming increase in unemployment. Second, the limited use of human resources — largely due to hierarchical dominance, strong seniority systems and the centralization of power — suppresses collective human thought and stifles human potential. Third, the mix-up of state laws and religious beliefs causes confusion and chaos through the misuse of religion's fundamental message about the ethical, moral and human ingredients of life. And fourth, there is an incoherent vision for science and technology.

The lack of a solid science and technology base is not always a result of poor capital or human resources. It sometimes stems from a lack of appreciation of the critical role of science and technology in development, an incoherent methodology for establishing such a base, and an absence of a coherent policy addressing national needs, and human and capital resources. Some countries consider scientific progress to be a luxury, in view of other demanding problems. Others believe that the base can be built by buying technology from developed countries. These beliefs

Aid programmes should be visionary in their mission and supportive of future investment.

translate into poor or, at most, modest advances that are based on the efforts of individuals, not institutional teamwork.

These issues point to three essential ingredients for progress. First is the building of human resources by eliminating illiteracy, ensuring active participation of women in society, and reforming education. Second is to rethink national constitutions, allowing for freedom of thought, minimizing bureaucracy, developing a merit system, and creating a credible — and enforceable — legal code. Third is the building of a science base.

The foundations of a science base are investment in special education for the gifted, the establishment of centres of excellence, and the chance to apply knowledge in the industrial and economic markets of the country and, eventually, the world. This must go hand-in-hand with a plan for general education at state schools and universities. With such a vision, a scientific culture will emerge that enhances a country's ability to follow and discuss complex problems, rationally and collectively. Scientific thinking becomes essential to the fabric of the society.

Many people feel that a scientific culture is only for developed countries. Some even believe in conspiracy theories — that the developed world will not help developing countries so as to control the flow of knowledge. I do not subscribe to such theories. The recent examples set by China and India, among others, of success in the world market result from their developed educational systems and technological skills in certain sectors. What is needed to develop a scientific culture successfully is acceptance of responsibility in a collaboration between developing and developed countries.

Developing countries need centres of excellence, not only for research and development, but also for training experts in advancing technologies and so reducing the brain drain experienced by many such countries. It is important that these centres are not just exercises in public relations: they should be limited to a few areas in order to build confidence and recognition.

To this end, national resources are needed to support research and development in a selective way, following well-established criteria based on merit and distinction. To guide national policy, government at the highest level should create an overseeing board for science and technology, formed from national and international experts. Without serious commitment to these principles, progress will remain limited.

Basis for success

Some developing countries have made admirable progress in these areas: and the development of India, South Korea and Taiwan reflect this. In Egypt, the University of Science and Technology is involved in an experiment being carried out on a 300-acre site given by the government on the outskirts of Cairo. The university awaits the approval of a new law that will make it a non-profit, non-government organization.

But it is also the responsibility of developed countries to focus aid programmes. Usually an aid package is distributed between many projects, with lack of follow-up leading to diffusion of resources and in some cases corruption, so the aid does not result in significant successes. Real focus can be achieved by establishing what I call 'partnership-guided aid', with a significant fraction of the aid being directed towards excellence using criteria established in developed countries.

There must also be a minimization of politics in aid. The use of an aid programme to help specific regimes or groups is a big mistake, as history has shown that it is in the best interests of the developed world to help the entire populations of developing countries. The aid programme should be visionary in its mission and supportive of investment in future developments. Developed nations either can give money as charity or they can become partners, providing expertise and a follow-up plan.

Some developed countries are recognizing the importance of partnership, especially with their neighbours. Examples include the United States and Mexico, and western and eastern Europe. Real success can be achieved provided there exists a sincere commitment to a partnership. This is in the best interests of both the developed and developing countries for peaceful coexistence in a world of civilized humanity. ■

Ahmed H. Zewail is in the Divisions of Chemistry and Chemical Engineering; and Physics, Mathematics and Astronomy, California Institute of Technology, Pasadena, California 91125, USA. This commentary is based on his address to the Pontifical Academy of Sciences at the Vatican.

Origins of inspiration

A four-dimensional view of two twentieth-century icons.

Einstein, Picasso: Space, Time, and the Beauty that Causes Havoc

by Arthur I. Miller

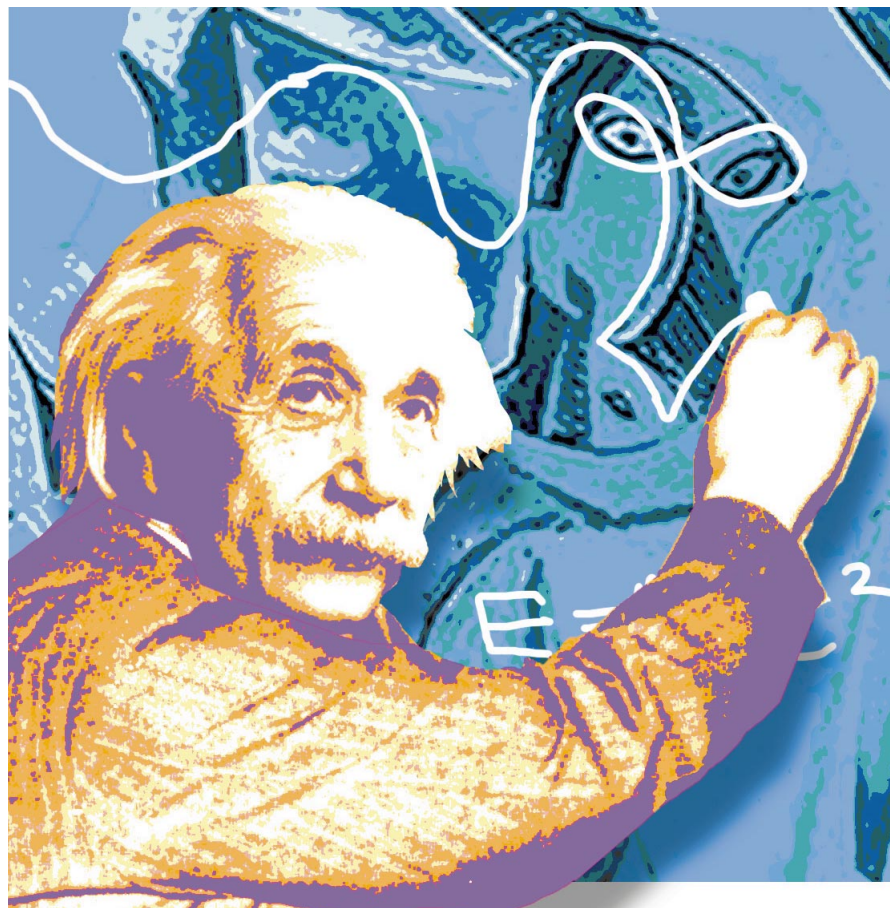
Basic Books: 2001. 320 pp. \$30

John Christie

The opening decade of the twentieth-century continues to hold a profound attraction for intellectual and cultural historians. It is not simply that obvious and major changes in painting, literature and music occurred at this time — the advent of modernism in Western culture — but that these changes seem so dramatically abrupt and, moreover, in some senses related to one another. The bolder souls among us also wish to include the sciences within this movement, so that the years 1905–09 witnessed a triple revolution at least, consisting of the departures initiated in painting by Picasso, Braque and Matisse in Paris, in music by Schönberg in Vienna, and in physics by Einstein in Bern.

Arthur Miller draws on his own substantial researches on the history of physics in this period to focus on two figures, Pablo Picasso and Albert Einstein. And, ultimately, on two key works — Picasso's *Les Femmes d'Alger* of 1907 and Einstein's 1905 epoch-making paper describing special relativity, "On the electrodynamics of moving bodies". Miller's account proposes a definite historical relationship between the work of his two creative heroes. He contends that both depended significantly on a preoccupation with geometry, and that this preoccupation had its basis in the work of the great French philosopher-mathematician Henri Poincaré. Miller is not implying that there was any communicative influence acting between Picasso and Einstein, but that their common preoccupations had an underlying common cause.

Miller provides fascinating and thorough accounts of the conception, gestation and delivery of the key works. He supplements them with two further types of material. Interested in personal creativity, he gives biographically based accounts of his protagonists. He also adds contextual detail on the cultural, philosophical and scientific environments that bore strongly on Picasso's and Einstein's developing ideas. So we are informed not only of the technical significance of ether-drift experiments and the Lorenz contraction, of primitivist and photographically derived aesthetics, but equally of the significance of the bohemian, hashish-devoted culture that existed in Paris's Montmartre, where Picasso spent much of this period, and of Einstein's and Picasso's love-lives. All this adds considerably to the book's readability and interest, and turns what could have been a rather dry and



testing disquisition on geometrical technicalities into a lively and often compelling narrative. It captures what Miller clearly feels to be the sheer, exhilarating excitement of those times, places and events.

Both Einstein and Picasso are, of course, massively archived and researched figures, and Miller's analysis of their achievements makes substantial and commendable use of the most recent research on Picasso — particularly that of Anne Baldassari on Picasso's photography, and John Richardson's biography. But Miller adds something significant of his own about the geometric origins and development of *Les Femmes d'Alger*. He identifies a plausible means by which the content of Poincaré's *La Science et l'Hypothèse* could have reached Picasso, who in those years did not have a strong grasp of the French language. Maurice Princet, a mathematically learned actuary, was a member of Picasso's circle of intimates, *la bande à Picasso*. And, according to Miller, it was Princet who told Picasso about the non-euclidean and four-dimensional geometrical speculations of Poincaré and also of the French mathematician Émile Borel in a way that stimulated Picasso's ambitions to apply them to his painting.

This is a valuable nugget of historical research, and is generally persuasive. If one additionally buys into Miller's unabashedly Romantic, genius-led views of creativity and historical change, then his book offers an account of Picasso in Paris that is both new and convincing, with the added bonus of having Einstein in the same interpretive frame.

It is possible, nonetheless, to accept Miller's factual findings while remaining sceptical about his overall interpretation. When questioned, Picasso persistently and over a long period denied a geometric or scientific inspiration for *Les Femmes d'Alger*. Miller's special pleading, that Picasso was not asked the question in the right way, is hardly convincing. It would be more to the point simply to insist that *Les Femmes d'Alger* is a geometrical composition: the presentation of space and objects in the painting are well described as geometrical, despite Picasso's disavowal. But a larger difficulty concerns the kind of geometry involved. It is very difficult to maintain, with Miller, that a Poincaréan 'four-dimensional' form is in play. Rather, three-dimensional depths, visually inaccessible in pictorial terms, are unfolded to the two-dimensional flatness of the picture plane,

giving the viewer the impression of seeing the picture from several perspectives, and thereby enhancing its visual effect. This is complemented by the fact that objects in the picture are geometrically broken up through the faceting that Picasso applied. Although it is interesting that contemporary occultists perceived four dimensions to the picture, this is difficult to validate from the point of view of the picture's composition. And it leaves open the question of how fundamental, or otherwise, is its derivation from Poincaré.

Perhaps surprisingly, the case for Einstein appears even more questionable. Although Poincaré's place in the prehistory of special relativity is long acknowledged, there is the stumbling block that Einstein's Special Theory is not presented geometrically by Einstein. He treated electrodynamics not in any markedly geometrical way but rather by an abstract reconsideration of the relations between rigid rods, perfect clocks and light signals — that is, by proposing a theory of measurement that respecified the nature and meaning of fundamental physical quantities.

Miller's argument may be on stronger ground at a later period — with the formulation of analytical cubism by Picasso and Georges Braque, and with general relativity's fundamental space-time geometry. But for the main period on which the book focuses, 1905–07, his case, however vigorously and enjoyably portrayed, remains tantalizingly unmade. ■

John Christie is at the Division of History and Philosophy of Science, University of Leeds, Leeds LS2 9JT, UK.

... but what does 'blue' smell like?

Synaesthesia: The Strangest Thing

by John Harrison

Oxford University Press: 2001. 250 pp.

£16.99, \$35

Ilya Farber

Of all the neurological disorders known to modern science, synaesthesia may be the only one that routinely inspires envy in those who study it experimentally. The experience of the synaesthetic seems somehow richer and more compelling; after all, who wouldn't want to experience music in full, glorious colour? Most 'normal' humans experience occasional cross-modal associations, perhaps with the assistance of Mahler or LSD, but this seems thin compared with the inner world of the synaesthetic painter who always experiences the spoken word 'Moscow' as "darkish grey, with spinach-green and pale blue in places". How can one not envy the woman who experiences shapes accompanying smells, so that the smell of wintergreen evokes "ragged



DAVID NEWTON

edges", and peanut butter smells like "things falling down and backward"?

For the past decade, John Harrison has been studying such people, and his book provides an accessible survey of what is known about this most intriguing of disorders. Although he wrote the book alone, most of the work he reports was done in collaboration with Simon Baron-Cohen, and the theory that they jointly developed forms the basis for much of Harrison's approach.

This theory begins with the well-established fact that all humans are born with substantial interconnections among the different sensory pathways in the brain. These normally die off within the first six months of life, but if a pathway's normal source of input is cut off, as occurs in congenital blindness, it may retain its connections to other pathways. Harrison draws the reasonable inference that in synaesthesia some of these pathways are somehow preserved into adulthood. He supports this hypothesis by pointing out that synaesthetics usually say they have had the condition for as long as they can remember, and that (according to an analysis of the limited pedigree data available) the most common form of the condition appears to be genetic — specifically, an X-linked dominant trait that is partially lethal in males.

Unfortunately, definitive tests for neuroanatomical connectivity are impossible in humans, as they require the use of injected tracers, followed immediately by post-mortem sectioning and staining of the brain. Harrison does present neuroimaging data

showing that, when colour-word synaesthetics listen to words while blindfolded, some of the brain areas activated are those that have been implicated in colour processing. But most of these areas are known to be driven by stimuli other than just colour even in normal people, and some of the areas activated in Harrison's imaging studies are not sensory at all. Even the clearest result, which shows activity in a highly specialized colour area, cannot identify the source of the activity. So it remains a possibility that the visual pathway is being driven not by cross-wiring but by something more like visual imagination.

Overall, Harrison's theoretical discussions are relatively superficial; where the book really shines is in discussing the data. Reading this book is like looking over a detective's shoulder at the collection of case-files and newspaper clippings spread out on his desk. Harrison has done all the arduous work of gathering the raw data and sorting through historical reports; now he invites the reader to join him in sifting through it for patterns and meaning.

And a fascinating deskful it is. Harrison calculates that the overall population prevalence of synaesthesia is at least 1 in 2,000; most of the synaesthetics he has found are word-colour synaesthetics, and about 85% of these are women — though, perhaps unsurprisingly, all of the famous historical candidates for synaesthesia that Harrison found are men.

It turns out that the colour of a word is usually determined by its first letter, and that

each synaesthetic's list of associations is idiosyncratic, with no discernible commonalities even among identical twins. But any one person's associations remain very stable over time. This is the basis of the now-standard 'test for genuineness', which checks the constancy of a subject's associations for a given list of words — non-synaesthetics tend to match fewer than 40% of their previous associations, whereas synaesthetics average more than 90%.

The organization of the book is somewhat frustrating. At times it's structured like an autobiographical narrative, at others it's more like a standard scientific monograph, while in many places it becomes an introductory text in scientific methods. Harrison's pedagogical digressions are sometimes important, but more often baffling. For example, the very brief discussion of neuroimaging data is preceded by more than 20 pages on the general history of neuroimaging. Ultimately, though, he doles out enough fascinating tidbits to keep the reader's interest, and the unanswered questions are tantalizing. I found the book especially good 'afternoon coffee-break' reading — although doubtless some synaesthetics would find that an unconscionable clash of tastes. ■

Ilya Farber is at the Philosophy-Neuroscience-Psychology Program, Washington University of St Louis, One Brookings Drive, St Louis, Missouri 63130, USA.

Creationism by stealth

Icons of Evolution: Science or Myth? Why Much of What We Teach About Evolution is Wrong

by Jonathan Wells

Regnery: 2000. 362 pp. \$27.95

Jerry A. Coyne

Opposition to evolution is found in many corners of the American religious landscape, including the Unification Church. Church founder Sun Myung Moon has frequently condemned darwinism for giving God no role in the history of life. In 1976, Jonathan Wells, a student in Moon's seminary, answered his leader's call. He writes, "Father's [Moon's] words, my studies, and my prayers convinced me that I should devote my life to destroying Darwinism, just as many of my fellow Unificationists had already devoted their lives to destroying Marxism. When Father chose me (along with about a dozen other seminary graduates) to enter a PhD program in 1978, I welcomed the opportunity to prepare myself for battle." The University of California supplied Wells with his weapon, a PhD in biology and, with *Icons of Evolution*, Wells has fired the

latest salvo in the eternal religious assault on Charles Darwin.

This personal history, taken from the Unification Church website (<http://www.tparents.org/Library/Unification/Talks/Wells/0-Toc.htm>), is conspicuously missing from the author's biography in *Icons*. The book, aimed at the non-specialist, masquerades as a scientific critique of classic examples of evolution, but is actually a polemic intelligently designed to please Father Moon. *Icons* is a work of stealth creationism, and strives to debunk darwinism using the familiar rhetoric of biblical creationists, including scientific quotations out of context, incomplete summaries of research, and muddled arguments. But because Wells has scientific credentials, studiously avoids mentioning religion or God (who appears only under the alias "intelligent design"), and presents his book as an objective critique (complete with 70 pages of references and research notes), it is easy for the non-scientist to be taken in. *Icons* has been embraced with glee by anti-evolutionists, who want it included in the American school science curriculum.

Wells's book rests entirely on a flawed syllogism: hence, textbooks illustrate evolution with examples; these examples are sometimes presented in incorrect or misleading ways; therefore evolution is a fiction. The second premise is not generally true, and even if it were, the conclusion would not follow. To compound the absurdity, Wells concludes that a cabal of evil scientists, "the Darwinian establishment", uses fraud and distortion to buttress the crumbling edifice of evolution. Wells's final chapter urges his readers to lobby the US government to eliminate research funding for evolutionary biology.

To see his argument at work, let's look at development, which Wells has referred to elsewhere as "the Achilles' heel of Darwinism". As Darwin first realized, some aspects of vertebrate development — especially transitory features — provide strong evidence for common ancestry and evolution. Embryos of different vertebrates tend to resemble one another in early stages, but diverge as development proceeds, with more closely related species diverging less widely. This conclusion has been supported by 150 years of research.

Wells tries to refute this mountain of work by noting that, in 1891, the German biologist Ernst Haeckel published illustrations of vertebrate embryos that exaggerated their similarity, and that some biology textbooks still display these doctored drawings. This embroidery, however, was first reported by the British zoologist Adam Sedgwick in 1894, and has repeatedly been used to show the failings not of darwinism, but of Haeckel (see, for example, *Nature* **410**, 144; 2001).

Despite Wells's arguments, one does not need Haeckel's wishful pencil to draw copious evidence for evolution from develop-

mental biology. Human embryos, for example, have pharyngeal pouches, a tail and six aortic arches — all features found in embryonic fish. But our pouches become glands and ducts instead of gill slits, our tail disappears, and our aortic arches (which remain six in some fish) either disappear or are transformed into carotid, systemic and pulmonary arteries. In our first trimester we develop the lanugo, a coat of hair that is shed before birth.

Are these patterns mere whims of the Intelligent Designer, or evidence of our common ancestry with fish and furry primates? Embryos of whales and some snakes develop hindlimb buds that regress before birth; embryos of baleen whales possess teeth that later disappear; and horse embryos have three well-developed toes, with the outer two shrinking to leave the single-digit hoof. Such examples abound, but you won't find them in *Icons*.

Wells also notes that the earliest vertebrate embryos (mere balls of cells) are often less similar to one another than they are at subsequent stages when they possess more complex features. According to Wells, this counts as evidence against biological evolution, which supposedly predicts that the similarities among groups will be strongest at the very first stages of development. But darwinism makes no such prediction. Darwin himself noted that embryos must adapt to the conditions of their existence, and the earliest stages of vertebrate embryos show adaptation to widely varying amounts of yolk in their eggs. Wells repeatedly fails to grasp the evidential value of phenomena that can be understood only as the result of a historical process, even if the results were not predictable. Perhaps an observer in the early Cenozoic could not have predicted that a lineage of ungulates would lose their hindlimbs as they became aquatic, but the development of the hindlimb in embryonic whales can be understood only as a result of descent with modification from a four-legged ancestor.

When discussing other 'icons', Wells uses the same tactic of selective omission to distort a body of literature he pretends to review. Nowhere is this more visible than in his chapter on human evolution. Faced with a series of hominid fossils showing transitions from ape-like to modern human traits over 4 million years, Wells can only mumble about the Piltdown Man hoax, and imply that the vigorous scientific debate about the course of human evolution proves that humans did not evolve.

It is telling that, although Wells repeatedly attacks evolution, he gives no hint of his own ideas about the origin and development of life. There is good reason for this. As one learns from his website sermon, *Evolution by Design*, Wells believes that "the human species was planned before life began, and that the history of life is the record of how this

plan was implemented". To Wells, the fossil record does not represent a continuum of ancestry, but a succession of creations by the Intelligent Designer, with each species carefully devised to nurture the next product of creation up to the final goal, humans.

But this argument is blasphemous, for its logical consequence is that the pinnacle of evolution is not *Homo sapiens* but our ectoparasite *Pediculus humanus*, the body louse. It also turns the Great Designer into a Great Deceiver, who, in the manner of Satan, put fossils in the rocks — and tails on embryos — to mislead biologists of the future.

Finally, Wells's main theme about the collusion of evolutionary biologists is simply wrong. Authors of some biology texts may occasionally be sloppy, or slow to incorporate new research, but they are not duplicitous. And, far from representing a conspiracy, it is invariably evolutionists (including myself) who have noted problems with some classic icons of evolution. Wells has it backwards. It is creationists like him who are conspiring to purge evolution from American education. They hide their own differences about issues such as the fossil record and the age of the Earth, they pretend to be disinterested seekers after truth, they fail to do their own scientific research, and, like Wells, they avoid at all costs revealing their own theories about the history of life. *Icons* is exactly as even-handed and intellectually honest as one would expect from someone whose "prayers convinced me that I should devote my life to destroying Darwinism". ■

Jerry A. Coyne is in the Department of Ecology and Evolution, University of Chicago, Chicago, Illinois 60637, USA.

A pillar of molecular biology

Ahead of the Curve: David Baltimore's Life in Science

by Shane Crotty

University of California Press: 2001. 372 pp. \$29.95, £19.95

Robert Bazell

Many of his peers regard David Baltimore as the most important living figure in biology. His co-discovery in 1970 of the enzyme reverse transcriptase, which makes DNA from an RNA template, opened the way for, among other things, recombinant DNA, the modern understanding of cancer and most of what is known about HIV. All too often, a scientist's significant research ends with such a seminal Nobel prizewinning discovery. But not only has Baltimore's productivity continued unabated, his work has left "a lasting impact on virtually every realm of modern biology", according to one colleague. His dis-

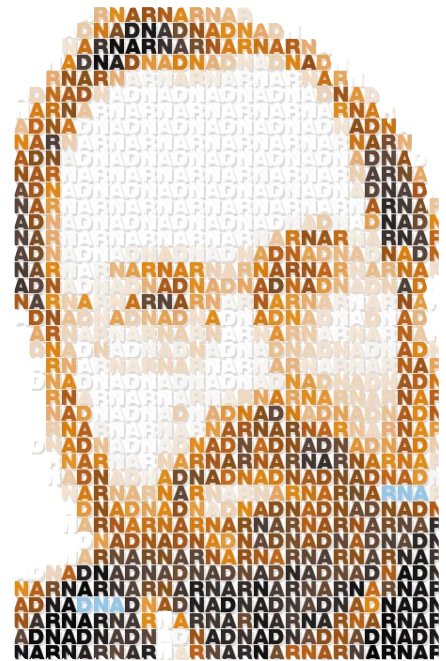
tinctive teaching style — a combination of frightening confrontation and passionate loyalty — has attracted and trained a large and elite corps of scientists who are now leaders in many areas.

And Baltimore's achievements extend beyond the research bench. He has served as president of two of the world's leading research universities: Rockefeller University, from which he was forced to resign, and the California Institute of Technology, where he now presides. He started and effectively ran the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts, one of the most productive and innovative models for public-private research cooperation. He was involved in most of the prominent science-policy issues of the past 30 years, including the US government's declaration of a 'war' on cancer, the initial perceived dangers of genetic engineering, the response to the AIDS epidemic and, of course, the most famous and protracted allegation of scientific fraud ever — the Imanishi-Kari affair or, as it is often called, 'the Baltimore case', in which the researcher Thereza Imanishi-Kari was accused of falsifying data for a paper co-authored with Baltimore and published in *Cell* in 1986.

Because of that case, or perhaps because some people use it as an excuse for other agendas, Baltimore — to put it mildly — faces detractors every bit as fervent as his supporters. James Watson, the co-discoverer of the structure of DNA and the embodiment of biology for the US public, actually campaigned to have Baltimore's Nobel prize rescinded and to have him expelled from the National Academy of Sciences. Others who were once considered Baltimore's close friends have behaved with only slightly less rancour.

Reading Shane Crotty's excellent book, we do not necessarily understand such horrid behaviour but we come to appreciate the context. Baltimore embodies molecular biology over the past 35 years, during which it evolved rapidly from an observational science to the Messiah of medicine, capable of delivering us from deadly degenerative diseases. It has become a magnet for so much money that even those marginally involved, not to mention the leaders, routinely make sizeable fortunes. This makes it fertile ground for monumental egos, and, much to the dismay of many of his contemporaries, Baltimore's powerful intellect and supreme, often abrasive self-confidence remained intact during the most trying times.

"Brilliant, eloquent and personable, Baltimore is a man whom even his closer friends refer to as arrogant and ruthless," Crotty, a postdoctoral fellow at the University of California, San Francisco, tells us early on. Unfortunately, we get little sense of the origins of Baltimore's personality. This is not a psychological biography. We learn that Balti-



DAVID NEWTON

more adored his mother, but she, along with his father, is not even named in the book. The mother pushed David into biology by suggesting a high-school summer programme at the Jackson Laboratories in Bar Harbor, Maine. Throughout his training, Baltimore seldom attached himself to one mentor for long, possessing a keen appreciation of those who could teach him something and great disdain for those who could not. From his earliest days as a graduate student he was "a pain in the ass", one teacher recalls.

But the reader will find few personal anecdotes here. At one point Crotty quotes Baltimore about his time as a graduate student at Rockefeller University: "I'd go to the movies occasionally, had some love affairs and things, but — I did an awful lot of lab work. Loved it!" That's as juicy as it gets. Baltimore's first marriage and divorce, and his subsequent marriage to the prominent virologist Alice Huang, merit only a few lines. Because of this lack of personal detail, the reader gets little insight into Baltimore's emotions during the many vicissitudes of his career.

What the book does offer is fine science writing. From Baltimore's first experiments with mouse genetics during his high-school summer to recent efforts to accelerate the development of an effective AIDS vaccine, Crotty describes clearly the underpinnings of each stage of Baltimore's career in language accessible to the educated layman, but not condescending to the practising scientist. As a result, we understand not only the research, but also the policy issues and personality conflicts it generated. By the time we are told about Baltimore's discovery of reverse transcriptase, we can thoroughly grasp its importance and the reason for the excitement. When Crotty details the experiences in Stanford biologist Paul Berg's lab with the animal tumour virus SV40, it is easy

to understand the initial terror Baltimore, Berg and others felt at the prospect of recombinant DNA experiments.

Crotty's retelling of the Imanishi-Kari affair also benefits from this clear detailing of the science. With the precisely focused lens of hindsight, and in enormous detail, Daniel Kelves' book *The Baltimore Case: A Trial of Politics, Science, and Character* (Norton, 1998) demonstrated the great injustice done to Imanishi-Kari and Baltimore. Crotty tells a more human story. The facts were not so clear at the time, and in the fog of misinformation Baltimore changed his own story more than once. This infuriated many who thought he went too far in attacking his critics, especially John Dingell, the bombastic Detroit congressman who chaired the congressional committee investigating the case. Although science itself was never on trial, as Baltimore and others alleged, Baltimore's ultimate exoneration was critically important for science because it restored one of its greatest practitioners to his deserved status. Crotty's book brilliantly illuminates this pillar of molecular biology, and should be read by anyone, whether scientist or not, who cares about the modern research enterprise and the politics that drive it. ■

Robert Bazell is at NBC-TV, 30 Rockefeller Plaza, New York, New York 10112, USA.

What every woman knows

Athena Unbound: The Advancement of Women in Science and Technology

by Henry Etzkowitz, Carol Kemelgor & Brian Uzzi

Cambridge University Press: 2000. 282 pp. £12.95, \$19.95

The Gender and Science Reader

edited by Muriel Lederman

& Ingrid Bartsch

Routledge: 2001. 505 pp. £55, \$90 (hbk), £18.99, \$29.95 (pbk)

Sylvie Coyaud

In spite of its subtitle, *Athena Unbound* has little 'advancement' to report in the lives of the American women scientists interviewed by the authors. Only the most determined were able to overcome early stereotypes of what girls should be like at home, at school or in society, as well as later prejudices. Girls tend to leave college with higher exam results than boys and with reasonable self-assurance. Then, from university onwards, they are made to feel out of place in labs, pubs and the other meeting places where gossip and information are swapped and where men receive their informal training in career-long networking.

The book contains few in-depth interviews, but the scientists selected have perceptive things to say about their own and other women's failures. Their testimony often contradicts prevailing assumptions and policies. For instance, they say that the impact of role models — older women in top jobs who function as beacons for young students — has been overrated. The poor 'models' have to struggle so hard to stay on top and are so obviously exhausted that only masochists would emulate them. The 'critical mass' theory, according to which a sufficiently large number of women researchers in a field would lead to a change in the power structure, doesn't stand the test either. This naive idea goes against experience: in some areas of the life sciences women are in the majority, but are kept firmly on the lower rungs, where a large and not too skilled labour force is needed. The same thing happened in medicine in European countries, and in road building in the Soviet Union.

The authors, two sociologists and a psychotherapist, want to dispel clichés about the inability of women to survive in a harsh environment. Not that they need to sell this argument — the large numbers of women raising a family as single mothers on a part-time salary are proof enough of their resilience. And although the authors' quantitative analysis method is respectable, job and pay statistics already tell the whole story. The book, though slim, feels over-long, because its unstartling conclusions are hammered home by the American habit of summarizing a chapter's contents at the end of that same chapter, as if readers were memory-impaired.

The Gender and Science Reader covers the same ground from the point of view of the feminist critique of science. It is a collection of essays and book extracts from the best English-language authors, from well-known

feminist writers such as Donna Haraway, Evelyn Fox Keller, Hilary Rose and Carolyn Merchant, to newer entrants such as biologists Christine Wennerås and Agnes Wold, who discovered the 2.6 factor (they showed that women had to publish 2.6 times more than men in order to obtain the same quality scores for postdoctoral fellowship applications submitted to the Swedish Medical Council; see *Nature* **386**, 341–343; 1997).

The book starts with a survey of (mainly US) data on the low status of women scientists. This is followed by a critique of scientific methods and values and of the underlying ideas about 'nature'. The editors nicely balance the different schools of feminist theory. For example, Sarah Harding's essay "Is science multicultural?" is followed by Helen Longino's "Subject, power, and knowledge", a review of feminist epistemological strategies. Longino herself favours "the inclusion of cognitive diversity", whether gendered or cultural, within the scientific community "as a resource for criticism of the received wisdom", but doesn't believe "that every alternative view is equally deserving of attention".

Longino is in turn followed by John Lukacs, a "historian of twentieth-century European culture who brings a strongly religious perspective to his discussion of... quantum mechanics". Lukacs treads less cautiously than the other two men who contribute to this anthology. One suspects his essay on "Heisenberg's recognitions", which discusses the "historicity of reality as something which is prior to its mathematicability" in relation to human nature, has been included to underscore the rationality of feminist authors. Comic relief is provided by chapters on "Gender practice" and "Science and identity", which tell how past research in the life sciences proceeded apparently unaware that human organisms, unlike bacteria, come in two versions, or



DAVID NEWTON

blinded by an eagerness to assign inferior quality to female cells, genes or brains.

The book's final section asks the question: will feminist scientists change science? Well, in some areas of science they already have, for instance in animal behaviour studies, Earth sciences and medical research. Hilary Rose, who has been chosen to provide an uplifting epilogue, sees "harbingers of hope". I agree. An increasing number of men love science enough not to deprive it of the "brilliance of women" (Rose). In the United States, where women complain more loudly than in most other countries, equal-opportunities policies have improved their lot in publicly funded research. But in Europe, according to "Women and Science", a report published last year by the European Technology Assessment Network, even countries committed by European treaties to fight discrimination in the work-place are reluctant to interfere with scientific bodies. If they don't interfere with teamsters' unions or soccer fans, who cares? But scientists take pride in their objectivity, yet don't notice that their sexist bias makes them look like the Victorians who preached family values to the underclass during the day and patronized brothels at night. No wonder so many women still think, with Virginia Woolf, that science "is a man, a father, and infected too". ■

Sylvie Coyaude is a science journalist based in Milan (e-mail: scoyaude@planet.it).



approach is of necessity non-mathematical, and many of the arguments proceed by analogy. In these cases it is Barrow's arguments that are the more convincing.

All three books tell the story of how zero was invented several times, finally impinging on Western mathematics from its incarnation in Indian and Islamic culture in time to spark the Renaissance. We can only join Pierre-Simon Laplace in acknowledging: "The ingenious method of expressing every possible number using a set of 10 symbols (each symbol having a place value and an absolute value) emerged in India. The idea seems so simple nowadays that its significance and profound importance are no longer appreciated. Its simplicity lies in the way it facilitated calculation and placed arithmetic foremost among useful inventions. The importance of this invention is more readily appreciated when one considers that it was beyond two of the greatest men of Antiquity: Archimedes and Apollonius."

Barrow includes an account of the parallel development of zero by the Mayan civilization of Central America and numbers his chapters with Mayan 'glyphs' (gargoyle-like stylistic faces used as numerical symbols).

The connection between zero and the nothingness at the heart of Barrow's book is made by way of that most slippery of concepts, the empty set. Nineteenth-century mathematicians were able to use some cunning tricks to make this most humble of sets the basis for their arithmetic. It is a curious fact that although we may worship different deities, we are all forced to use the same

empty set. It is the true universal concept. No wonder that this idea, which mathematicians now accept with equanimity, caused much debate and heart-searching down the centuries among philosophers and theologians.

Early scientists, despite the reservations of the philosophers, pursued the vacuum increasingly successfully with their barometers and pumps, and the idea of an empty space seemed eminently satisfactory. But then the necessity of finding a medium to allow the propagation of various electromagnetic waves meant that theoreticians were led to fill it up again with a presumed 'ether' — an invisible, inviscid fluid permeating all space. Physicists played intricate games with this new toy. Lord Kelvin and Peter Tait knotted ethereal vortices; Simon Newcomb and John Gore invoked its properties to explain the darkness of the night sky. Theologians likened it to "The Rudder of the Universe, the Rod ... in the Hand of the Almighty".

With the benefit of hindsight, we can see that by the end of the nineteenth century the theory was running into trouble. This was the beginning of big-money science, and it was the first telecom magnate, Alexander Graham Bell, who financed the experiments that laid the ether to rest. Albert Michelson and Edward Morley found strong evidence for its non-existence, and eventually the man-in-the-street's favourite eccentric genius, Albert Einstein, devised a theory that allowed it to be dispensed with altogether and so restored the status quo for an empty vacuum. It was a pity he had to use mathematics that the man-in-the-street could not possibly understand.

Nature may not always abhor a vacuum, but theoretical physicists seem unable to leave it alone. Only a short time after Einstein had restored empty space as a concept that one could believe in, the invention of quantum mechanics made it something that was too simple to satisfy the new theories. In the most persuasive model, any so-called empty box contains innumerable pairs of virtual particles which switch themselves in and out of existence, and which contribute to the energy of that apparently empty hole. This is the zero-point or vacuum energy. Even the experimentalists joined in, and measured the tiny forces involved in the Casimir effect, one of its consequences. Barrow brings this firmly into the realms of credibility by outlining results on a nautical phenomenon noticed in the nineteenth century, when two ships wallowing near each other in a choppy sea were moved towards each other by the wave action. It turns out that this involves the same theory, albeit on a much greater scale.

Such is the way that modern physics is interconnected that these effects noticed in the 'quantum vacuum' turn out to have wide-ranging consequences even at the truly astronomical scale. Cosmologists now have to take account of this new version of empty

Nothing to it!

The Book of Nothing

by John D. Barrow

Pantheon: 2001. 370 pp. \$27.50

John O'Connor

It must have been all those noughts at the end of last year's date that started people writing about zero. In 1999 we had Robert Kaplan's *The Nothing That Is: A Natural History of Zero* (Penguin), in 2000 we had Charles Seife's *Zero: The Biography of a Dangerous Idea* (Souvenir), and now we have John Barrow's *The Book of Nothing*. It's like standing at a bus stop. For 1,000 years nothing arrives and then you get three books on zero coming along together.

Kaplan's book was content to examine the history of zero and its position in a mathematical framework. The other two take the exploration of the idea out of the domain of pure mathematics and into the modelling of the physical Universe. This makes for a livelier story and expands the interest of the topic away from the purely arithmetic. Of these two, Barrow's book is the more authoritative and tells the story more persuasively. As with all popular books on modern physics, the

space. Theories of how the Universe evolved in the first few moments after the Big Bang can be profoundly influenced by this concept of energetic nothingness.

In 1992, John Barrow wrote *Theories of Everything: The Quest for the Ultimate Explanation* (Clarendon). It is curious that a decade later the ultimate questions can be explored from the point of view of nothingness rather than everything. ■

John O'Connor is in the School of Mathematics and Statistics, University of St Andrews, St Andrews KY16 9SS, UK.

Dispelling the boredom

Stimulating Concepts in Chemistry

edited by Fritz Vögtle, J. Fraser Stoddart & Masakatsu Shibasaki
Wiley-VCH: 2001. 414 pp. \$50, £29.95

The New Chemistry: A Showcase for Modern Chemistry and its Applications

edited by Nina Hall
Cambridge University Press: 2000. 493 pp.
£30, \$49.95

G. M. Whitesides

Each discipline in natural science has a popular reputation. Biology concerns itself with life and is intrinsically fascinating; physics does something very difficult and unfamiliar; astronomy looks at far-away objects and provides amazing pictures; chemistry makes paint and is boring. For those of us who know and love chemistry, this last view is a painful one. Our opinion is quite the opposite: that everything that one touches or sees or tastes is chemistry — that, in fact, all of perceptible reality involves conversations among atoms and molecules — and that an appreciation of chemistry is an essential and endlessly engaging component of an appreciation of the world.

Occasionally, groups of us gather our strength and try to explain why it is that molecules, and their behaviours and idiosyncrasies, are so fascinating. These two books are the fruits of such endeavours. Although both books are efforts to 'explain' chemistry, they have different scopes and styles. Both succeed — in the style of many multi-authored collections — by assembling a collage rather than painting a picture. In looking at a collage, of course, what you see depends on how far back you stand. And in a collage, although the edges can be rough and colours missing, there is often a sense of vigour. Both books convey that sense of vigour for chemistry.

Stimulating Concepts in Chemistry is exclusively focused on organic chemistry and its borders with neighbouring fields, especially biology, materials science and

chemical engineering. Its chapters are in the style of reviews in technical journals, rather than essays intended to give a broader view. It is not concerned with the philosophy of its subjects — more with their nuts and bolts.

Organic chemistry is currently in a phase of change and expansion. For a number of decades, it has been focused on the art of organic synthesis. Its accomplishments in developing synthetic pathways to natural products — the often staggeringly complex molecules that are the odd offspring of metabolism — have been some of the great triumphs of the discipline. The utilitarian rationale for this area of research was that it developed methods that made possible the synthesis of drugs, as indeed it did. It was also intellectually very interesting — a grand game of strategy and logistics.

It is striking that very little of this classical genre is represented in these chapters. Their authors instead represent four different groups striking out across the marches. One develops synthetic methods — that is, refines important classes of reactions or processes — and traffics across the busy border between chemistry and chemical engineering. A second designs and fabricates organic molecules possessing new kinds of function: as wires for conducting electrons, molecular-scale machines, antennae for collecting light or molecular sponges that selectively adsorb other molecules. The authors of these chapters are collectively in the vanguard of those who believe that chemistry is a natural centre for nanoscience, as molecules are, in fact, exquisitely engineered and eminently manufacturable nanostructures.

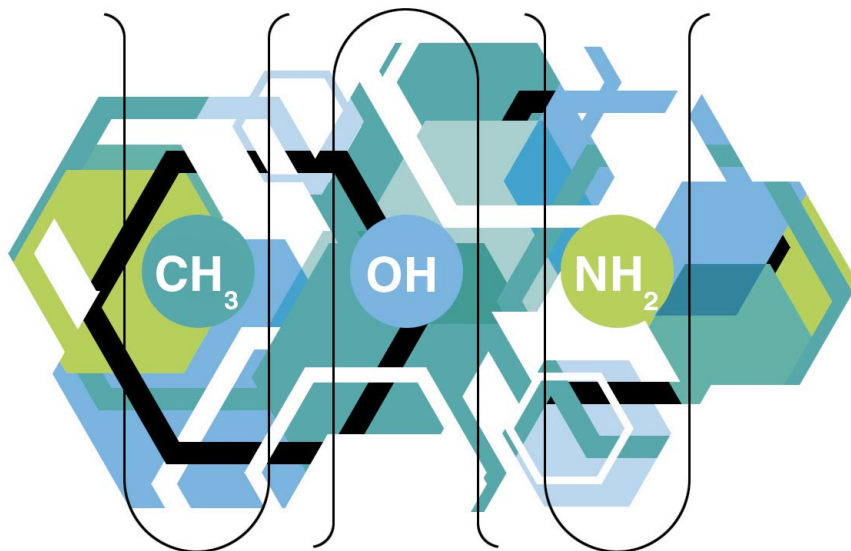
A third group focuses on the synthesis of molecules relevant to biology. The molecules they describe mimic, inhibit or modify enzymes — all of these are activities that will be increasingly important as biology turns from cataloguing the genome to understanding the proteins whose structures it encodes. A fourth group simply makes new

structures — exotic constructions containing only carbon atoms, multiply branched and rebranched molecules called dendrimers, and molecular crystals with designed structures. Chemists have always enjoyed making new molecules, and these chapters are baby pictures of new members of the molecular family.

The New Chemistry is a quite different book, and is intended, I think, for a more centrist audience. It is more classical in its coverage: it has a broader scope within chemistry and includes more history, but is less concerned with trade at the borders between chemistry and other disciplines. Its contributions span the full range of atomic and molecular behaviour: from the search for new elements to the production of energy, and from discussions of the nature of metals to discussions of chaos. It is surprisingly unconcerned with the intersection of chemistry and biology.

Is its title, *New Chemistry*, an accurate description? Well, yes and no. The fact that the essays cover a wide range of topics suggests, accurately, that the traditional subfields of chemistry are amalgamating. Some of the subjects — the chemistry of compounds held together by weak physical interactions rather than by covalent bonds, molecular electronics, materials, the production of energy, processes far from equilibrium, surfaces — are probably all, for different reasons, in for long runs. Others, such as the synthesis of complex organic structures, the study of the chemical bond and the chemistry of inorganic ions, are so fundamental that they will certainly always be part of the field, but may not be where most of the students go. The virtual absence of contributions from areas such as biochemistry and environmental chemistry reflects the fact that every collage is made with the materials available, and some colours and textures are always missing.

So neither book provides a broadly accessible, poetic view of chemistry. Both require



that the reader understands the periodic table and can interpret organic structural formulae. I will give my graduate students the reviews collected by Vögtle, Stoddart and Shibasaki. They provide an excellent, if idiosyncratic, view of new directions in organic chemistry — a field that has become actively evangelical, and is spreading the gospel of organic synthesis and synthetic nanostructures to all who will listen. Hall's book I will keep for myself. It covers a sufficiently broad scope that even those who have spent too many years thinking about atoms and molecules will find that its essays contain surprises. I will not trouble my wife with either: neither book is really suitable for explaining chemistry to those who are not chemists, although much of Hall's book should be easily understandable by other natural scientists.

Both books do a good job of describing chemistry to chemists in ways that catch the imagination. That is a good start in escaping the dread adjective 'boring'.

G. M. Whitesides is in the Department of Chemistry, Harvard University, 12 Oxford Street, Cambridge, Massachusetts 02138-2902, USA.

A constructive scheme unravelled?

The Collapse of the Kyoto Protocol and the Struggle to Slow Global Warming

by David Victor

Council on Foreign Relations/Princeton University Press: 2001. 169 pp. \$19.95, £12.95

Michael Grubb

The Collapse of the Kyoto Protocol is a serious, detailed, but ultimately paradoxical book. Apart from the introductory chapter — which contains some misleading caricatures about the Kyoto Protocol, the international agreement that set targets for greenhouse-gas emissions — it is written with an authority and detail that few can muster. It is also an absolutely independent assessment, beholden to no one. Yet the core conclusion reveals a fundamental inconsistency in the analysis.

David Victor gives short shrift to many of the sacred cows of US opposition to the protocol. Whereas most US criticisms centre on the lack of quantified commitments for developing countries, Victor argues the opposite, that the commitments should have focused only on the rich countries of the OECD. It would then have avoided the institutional difficulties raised particularly by including Russia and Ukraine. Most US economists argue that Kyoto's targets are too close, but Victor says they should have been closer, to minimize the uncertainties. Kyoto's targets cover many different gases; whereas opponents want the effort to focus more on



gases other than carbon dioxide, Victor says the targets should have been restricted to carbon dioxide only, because it is hard to monitor the others properly.

The core of Victor's analysis is thus an institutional critique, particularly with respect to emissions trading — the option for countries to trade their emissions allowances. Trading, he rightly asserts, is an inevitable consequence of adopting emissions targets. Yet he believes that bringing money into emissions control will create incentives to cheat the system that will exacerbate the already difficult problems of monitoring and enforcement. He also claims that, although Kyoto's flexibilities would allow the United States to meet its commitment through investing in emission reduction projects or systems abroad, the foreign institutions concerned may not be able to guarantee that this will really reduce emissions elsewhere.

Victor's book is built on years of experience and commentary on the climate-change regime. He has always been a critic of the targets-and-trading approach, and sees the collapse of the recent talks in The Hague as vindicating his stance. But his final chapter marks an important evolution in his thinking. Previously, he argued for a soft-law approach based on national reporting and review — a climate equivalent of the OECD's economic review process. Here he acknowledges that something stronger is needed. The answer? A mix of targets-and-trading with a tax system that sets a limit on the market price of emission permits.

Putting a cap on permit prices and restricting the system to carbon dioxide in the OECD countries, he argues, would overcome the institutional difficulties he has identified. But one senses an air of desperation. By the end of the book Victor has argued himself into supporting many of the Kyoto Protocol's fundamental principles. The problems he identifies are not unmanageable, and the changes he seeks are relatively minor compared with the prospect of negotiating an entirely new system. But as a long-standing critic, he cannot bear to recognize that. He would rather tear down the house and start again according to his own specifications.

Despite all its important insights, its ultimate negativity — and political naiveté — makes this a depressing book. For example, the difficulty of monitoring methane emissions does not preclude keeping the Kyoto basket of gases for national targets, while domestic policies target carbon dioxide and other gases separately. Monitoring need not involve unacceptably intrusive international inspection; a reasonable degree of democracy and transparency could minimize the scope for abuse. Victor complains that no one compared the benefits of including methane against the costs of (imperfect) monitoring, but he makes no attempt to compare such costs with his own alternative of abandoning the protocol along with the whole idea of cost-effectiveness across the different gases.

Victor's analytical strength and peculiar negativity are nowhere more striking than in his treatment of liability. He knows that emissions trading can do much to solve the thorny problem of compliance, provided that the validity of acquired emission allowances is made dependent on the selling country complying with its commitments ('buyer liability'). The alternative of seller liability "invites disaster ... it will fail because it is a fiction in international law". With buyer liability, however, selling countries would not get a good price unless they could convince the countries interested in buying that they have both the systems and the intent to comply. Indeed, the case Victor presents for such buyer liability is unanswerable and unsurpassed in the whole of the literature. Yet he then meekly notes that the "international debate" (actually the US position) at The Hague conference supported seller liability. He could have castigated the previous US administration for its cowardice and lack of imagination on the issue, and argued passionately that the rescheduling of the suspended Hague talks offers the world a chance to get this cornerstone of international emissions trading on the right footing. Instead, he simply concludes that the protocol is hopeless, but that if the world starts again with his own scheme, it will get the liability rules right.

There are too many such inconsistencies between Victor's critique of trading in the protocol and his endorsement of trading

according to his scheme. This is coupled with the casual assumption that the Kyoto deal can be completely rewritten. Victor vastly underestimates Kyoto's achievement in obtaining agreement and the centrality of the strong US target in return for extensive flexibilities. Indeed, despite its analytical depth, the book is marred — and perhaps the negativity explained — by viewing everything through a US lens. This results in scepticism about US implementation and cynicism about the rest of the world. For example, it is grossly misleading to assert that "no major government has a viable plan for compliance". Ten years in advance of the Kyoto target period (2008–12), the United Kingdom is on track, and, like many other European governments, if its current plans prove inadequate, they will be strengthened and/or supplemented by using Kyoto's mechanisms to achieve compliance.

There are other peculiar inconsistencies. At the core of his attack on emissions trading is the assertion that negotiating allocations is the equivalent of agreeing on how to distribute money, and is hence impossible, or too weak. Yet Kyoto did reach agreement — Victor's claim that all this was an accident born of governmental ignorance is arrogant and does not hold water. And his complaint about Kyoto is that the targets are too strong. Indeed, in the United Nations, the European Union, and even in individual governments, negotiations regularly distribute budgets. All that is needed is the common understanding that there is a collective interest in agreeing, and that simply printing more money only leads to inflation. Victor has not factored in the politics of pressure in the hothouse of negotiating allocations. It is the achievement of having agreed in Kyoto that Victor wants to unravel.

Ultimately, Victor's book reminds me of Tolkien's Lord Denethor in *The Lord of the Rings*, who was given a crystal ball to see events around the world — but only negative ones. Overwhelmed with despair, he led his family to the pyre, decrying the ignorant fools who fought on. He burnt to death just as his allies were mustering for their final victory. David Victor's analysis is that of a fine mind overwhelmed by the problems he sees and led to a destructive conclusion.

Nevertheless, the book should be read by all policy-makers attempting to find international solutions to the climate problem. It identifies the important questions to be answered, and the big problems to be solved. There are answers, and ironically, some may be easier to find while the present US administration is withdrawn from the Kyoto negotiations. Do not follow Victor onto the pyre. ■

Michael Grubb is at Imperial College, London SW7 2BP, UK. He is author of *The Kyoto Protocol: A Guide and Assessment* (RITA/Earthscan, London; Brookings, Washington).

Knowing the value of nature

Wild Solutions: How Biodiversity is Money in the Bank

by Andrew Beattie & Paul R. Ehrlich
Yale University Press: 2001. 272 pp. \$25.95

E. J. Milner-Gulland

Every textbook on conservation contains a section where the authors attempt to justify why policy-makers should make the conservation of biodiversity a priority. Although people tend to become conservationists for reasons more to do with the heart than the head, those wishing to make a strong case for conservation tend to downplay arguments based on ethics, aesthetics or spiritual needs, and concentrate on the utilitarian reasons for conservation. It is argued that only if biodiversity is seen to be paying its way, preferably in the short term, will funding for its conservation be forthcoming.

Wild Solutions argues for the conservation of natural resources from a firmly utilitarian perspective. Taking examples from a wide range of taxonomic groups, Andrew Beattie and Paul Ehrlich demonstrate how often-overlooked species are fundamental to our continuing viability. There are many fascinating examples of useful species, such as the glue-squirting velvet worm and the antibiotic-producing Australian bull ant. The authors cover a range of topics, including decomposition, medicines, problems in construction and motion, water and nutrient cycles, and bioindicators.

Using the concept of the 'natural internet', they describe how the natural world is composed of connections, unknown to most people, which are allowing us to dispose of

our waste, clean our air and water supplies and pollinate our crop plants. Along the way, the authors also provide powerful ammunition for those who argue that taxonomic studies are crucial to our future well-being. They show how little is known about the planet's biodiversity, and give examples of obscure creatures, stumbled upon by chance, which were subsequently found to be useful to humanity.

The book is almost unrelentingly positive, and an easy read, generating a momentum of energy and excitement about the potential of the natural world to solve many of the problems that face us. By finding where a problem that we face occurs in the natural world, they argue, we are likely to find solutions, either as blueprints for technology or in the form of inspiration. For example, termite air-conditioning has been used as the inspiration for naturally cooled office blocks.

Generally, though, the examples given are discussed briefly, and in little depth — this is an inspirational book for the general reader rather than the scientist. But the chapter on biological control does touch on the complexities involved in exploiting 'wild solutions' to our problems. It explains how introducing new species to solve one problem can create another — the cane toad, water hyacinth and *Salvinia* are among the examples given.

Reading the book from a UK perspective, however, I was concerned at the absence of any exploration of the issues surrounding the genetic modification of organisms. These are complex issues, hard to discuss in a book of this type. Nevertheless, the target audience in the United Kingdom must be convinced of the benefits of gene technology before it will accept the authors' premise that a combination of sophisticated biotechnology and naturally produced chemicals is a good way to proceed.

The rise of ecological economics as a discipline has made many conservationists believe that, in order to put their case across, they must value biodiversity monetarily. This confuses utilitarian value with monetary value. Beattie and Ehrlich do not attempt this, which I find enormously refreshing. In fact, a strong implicit message in the book is how complex and poorly understood are the mechanisms behind ecological processes, and how meaningless it would be to attempt a full monetary valuation of them.

Some impressive examples of the monetary value of natural services are cited, however. One is the Catskill Mountains, the source of New York city's water supply. The increasing load of pesticides, fertilizers and sewage had threatened to overwhelm the natural water-purification services provided by the Catskills. The costs of constructing a water treatment plant were



estimated at \$6–8 billion, with annual operating costs of \$300 million. By comparison, reducing the pressures on the ecosystem, allowing it once more to provide high-quality water, was estimated to cost \$1.5 billion. And this was the course of action chosen.

But as the authors say, "... our most important capital is natural capital: biodiversity ... *this* is the capital of the real world. The currencies based on financial capital derive from the life-support systems and products generated by biodiversity. We know full well that humanity has chosen to operate in a world of financial capital. The way to proceed, therefore, is first to recognize the true worth of the two kinds of capital and then to organize the human economy to preserve both. The value of financial capital is accepted by most people; the value of natural capital is only starting to be recognized."

This book has a limited remit: to demonstrate to the lay reader that the natural world is already supporting our species in ways unknown to many of us, and that many other useful products and services await discovery if we look for them in the right way. The destruction of biodiversity is damaging the life-support systems on which we depend, and cutting off our options for the future. By using a huge range of examples, Beattie and Ehrlich get the point across very effectively. ■

E. J. Milner-Gulland is at the T. H. Huxley School of Environment, Earth Science and Engineering, Imperial College London, London SW7 2BP, UK.

Murder most putrid

Maggots, Murder and Men: Memories and Reflections of a Forensic Entomologist

by Z. Erzinçioğlu
Harley Books: 2000. 256 pp. £13.95

Mark Benecke

Here is that rare thing, a good popular book on forensic entomology that is also an illuminating read on forensic science itself and on the art of being an expert witness. Zakaria Erzinçioğlu (a.k.a. Dr Zak), a forensic entomologist for more than 20 years, covers not only the wonderful world of insects as a tool in forensic investigations, but also the Tertiary geological period, O'nyong-nyong disease, Napoleon Bonaparte, human behaviour, maggot therapy and Sherlock Holmes.

Starting from the statement that "a dead human body is a magnificent and highly nutritious resource", mostly for invertebrates, Erzinçioğlu sets off on a fascinating trail that takes in both forensic entomology and the nature of forensic evidence. The insect evidence in two of his cases perfectly illustrates the fact that, although much can be ignored at a crime scene, it is important not to overlook vital evidence, or even its absence. In one case, the presence of winter gnats at the crime scene was a plain indication that the victim had died in winter. In another case, however, the presence of minute black scav-

enger flies (which start to breed earlier in the year than other flies) could not be interpreted correctly without recognizing the *absence* of blowfly maggots. If death had occurred in summer, blowfly maggots would have out-competed the tiny scavenger flies. The victim must therefore have died in the spring, when blowflies are not active.

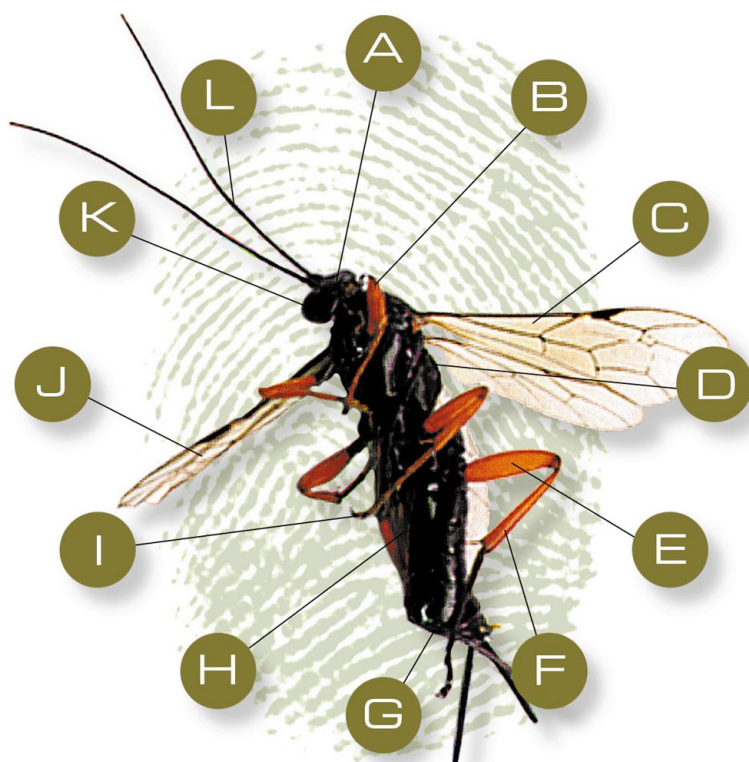
In contrast to the everyday work of most natural scientists, forensic scientists, especially the few who work at crime scenes, have to accept that controlled experimental conditions must frequently remain a dream. The complexity of influences in real-life situations can only be simulated by a large number of laboratory experiments, for which there may be no time. To highlight this problem, Erzinçioğlu discusses the scientific method itself. Because physical evidence from a crime scene can often only be interpreted, instead of delivering an obvious absolute truth, Erzinçioğlu warns against the danger of going too far. To avoid this trap, a forensic expert not only needs experience but must also have a clear idea of which questions can be answered by a particular investigative method and which cannot.

Case-work can be thrilling, but is also a personal challenge. As Erzinçioğlu authentically describes, one has "to visit the seediest of dwellings at the most unsocial of hours", carry out "post-mortem examinations late at night" and be "ferociously cross-examined by hostile barristers in court". He shares with the reader the mind-broadening knowledge he has gathered from working in such conditions, and which is the forensic scientist's reward.

Only a mixture of deep specialist knowledge (in Erzinçioğlu's case, the biology and anatomy of flies and their maggots), lots of enthusiasm and curiosity, up-to-date knowledge of neighbouring disciplines and a touch of singularity can make a scientist into a 'real forensic', and Erzinçioğlu is a prime example. In his discussions of the nature of justice he has already ruffled some feathers, for he points clearly to unpleasant and often neglected aspects of the justice system, such as incompetent defence consultants who muddy the waters and the fact that forensics in Britain apparently now has to operate according to 'free market' rules.

The energy and efforts of scientists such as Erzinçioğlu have put forensic entomology back into routine use — as long as a forensic entomologist is available. A special issue of *Forensic Science International* on forensic entomology is due later this year. So buy this book and enjoy the author's witty, wise and comprehensive thoughts on forensic entomology and forensics in general. Also, try not to steal the lovely black-and-white drawings by Michael Roberts as slides for your future forensic presentations — if you can. ■

Mark Benecke is Director of International Forensic Research & Consulting, Postfach 250411, 50520 Cologne, Germany.



Being objective

The idea of scientists as impartial observers is hard to shake, but is complete detachment justified?

Mary Midgley

What does it mean to be objective? John B. Watson, the founding father of behaviourism, had no doubts about this. For him, the word meant simply 'avoiding emotion'. Thus, in giving advice about parental behaviour, he wrote:

"There is a sensible way of treating children. Treat them as though they were young adults... Let your behaviour always be objective and kindly firm. Never hug and kiss them, never let them sit on your lap. If you must, kiss them once on the forehead when they say good night. Shake hands with them in the morning..." (*Psychological Care of Infant and Child* 9–10; W. W. Norton, New York, 1928).

As he knew that parents were prone to make this kind of mistake, he had grave doubts about whether they should be allowed to bring up their children at all, or indeed to have anything to do with them. "There are," he noted, "undoubtedly more scientific ways of bringing up children, which probably mean finer and happier children."

Here, as often happens, the notion of *objectivity* seems to be closely linked to that of *science*. But just what does 'scientific' mean here? It does not mean that these new methods had been scientifically proved to have this better effect, for they had not. Watson seems to mean by it something more like 'ways that avoid emotion, as a scientist may be expected to do when he (repeat he) is investigating something inanimate'. By objectivity, similarly, he means treating what one has to deal with as a lifeless object, not as a subject.

But can this precaution always be observed when that item is in fact alive, and may even be human? This question has worried many theorists. Some have dealt with it in heroic style, by ruling that life itself is actually an illusion. Thus Peter Atkins: "Inanimate things are innately simple. That is one more step along the path to the view that animate things, being innately inanimate, are innately simple too." (*The Creation* 113; W. H. Freeman, New York, 1987.) Similarly, Jacques Monod laid down his 'postulate of objectivity' — the assumption that no such thing as purpose is to be found anywhere in nature — and considered it "consubstantial with science", a universal law because, as he rather oddly put it, "Nature is objective" (*Chance and Necessity* 30, 54; Collins, Glasgow, 1972).

Now Monod, being a molecular biolo-

Life, consciousness and purpose are natural facts in the world like any others, and in some contexts they are vitally important facts.

gist, never had to consider whether this law actually extended to cover everything in nature, including active organisms such as small children. And again — as he too was part of nature — was his own purpose in writing his books real or not? He doesn't say. He took all purpose to be illusory and the difference between living and lifeless matter, sentience and unconsciousness, to be somehow a trivial one, beneath the notice of science. For this strange metaphysical view he claimed the authority of Galileo, as Galileo had excluded purpose from physics. But physics, as we know, is not the whole of science. It never has to deal with living subjects.

This strange notion of objectivity does not follow from the ordinary use of the word, which is simply *fair, unbiased, impartial*. In that sense, the objective way to tackle a question — for instance, whether it is good for small children to be hugged — is to try to find out the answer by studying the children themselves rather than by simply expressing one's own emotional response to this phenomenon, which is what Watson clearly did.

Life, consciousness and purpose are natural facts in the world like any others, and in some contexts they are vitally important facts. That is why we have to take subjectivity seriously. But the idea that our scientific duty is to treat everything we study as an inanimate object has been strangely persistent, long surviving the behaviourist ideology out of which it grew. When Jane Goodall first submitted an article about chimpanzees to *Nature*, the editor crossed out the words 'he' and 'she' where they were applied to the apes and replaced them by 'it'. Similarly, her use of names rather than numbers for the animals caused alarm, although psychologists dealing with human beings had established before this that using names was a much more efficient method for observing people than using numbers.

Is this greater complexity in living sub-



Lessons in objectivity: should parents, or even scientists, always be objective?

jects somehow damaging to the kind of objectivity that matters? Certainly it is harder to be impartial in studying children than it is in studying rocks. But we have to deal with this difficulty by watching for the obvious sources of bias, not by suppressing our natural responses altogether. Someone who sees children merely as unconscious physical objects is not going to be able to observe them very intelligently.

It is interesting here to note the impressive advances in primatology that have been made by Jane Goodall and the crowd of distinguished women scientists who have followed her — Dian Fossey, Birute Galdikas, Sarah Hrdy, Sue Savage-Rumbaugh and the rest. Although they are thoroughly rigorous in their methods, what has drawn these people to this particular study is their deep sympathy and respect for the animals that they investigate. That sympathy and that respect lie at the core of their science, as well as fuelling their strong efforts to save these species from extinction. A watsonian fear of emotion that ignores such motives overlooks something central to the scientific spirit. ■

Mary Midgley was formerly senior lecturer in philosophy at the University of Newcastle upon Tyne.

Moving rhythms

Eve Marder

One hundred years ago, neuroscientists watched children and cats walking, jumping and running, and realized that these and other rhythmic movements require the orderly and progressive activation of many different muscle groups. How does the nervous system achieve alternate contraction of extensor and flexor muscles, timed correctly so that the skipping child does not trip or the flying locust fall from the sky?

On first principles, you, like the early neuroscientists, might propose two different classes of models to explain the generation of rhythmic movements. First, you might suggest that sensory receptors would continuously monitor the position of all limbs and joints, thus activating reflexes to trigger the next appropriate movement. Second, you might suggest the existence of an oscillatory circuit in the brain stem or spinal cord in vertebrates, or in central ganglia in invertebrates, to provide the essential timing for the sequential activation of the muscles needed to produce the desired repetitive movement. The former mechanism has been termed 'chain of reflexes' and the latter 'central pattern generator'.

Many neuroscience textbooks and almost all general biology books introduce the myotactic or stretch reflex. Consequent-

ly, it is often mistakenly assumed that the stretch reflex is the basis of rhythmic movement. For historical reasons, biology books and many neuroscience courses give little or no treatment to the large body of experimental work showing that rhythmic movements are organized by oscillatory central pattern-generating circuits in invertebrates and vertebrates.

Today, we define central pattern generators as circuits that can produce rhythmic motor patterns in the absence of phasic drive, either from a sensory input or a descending control pathway. In other words, a central pattern generator is present when the existence of a rhythm does not depend on specific timing cues from peripheral or higher centres in the animal, although these inputs will certainly alter the motor patterns produced by the central pattern generator.

The existence of central pattern-generating circuits was first demonstrated by experiments showing that rhythmic movements are still produced after all the sensory feedback from limbs is removed. Today, we study 'fictive' motor patterns produced by pieces of the nervous system removed from the animal, similar to those that would drive movement if the muscles were attached, but termed fictive because the nervous system is no longer attached to its target muscles and sensory inputs. In many invertebrate animals, isolated ganglia produce fictive motor patterns. Spinal cords isolated from rats, mice, lamprey and chicks are all able to generate them, and slices of vertebrate brain stem generate fictive respiratory rhythms.

How does a group of neurons produce oscillations driving different motor neurons to fire rhythmically out of phase with one another so that the left and right legs can step in alternation? There are two general strategies for producing rhythms. Some are driven by neurons that are themselves oscillators, and that act as pacemakers. These neurons are similar to the pacemakers in the heart, and have specific combinations of voltage-dependent ion channels causing them to fire rhythmically in bursts of action potentials in the absence of rhythmic synaptic drive. It is thought that the circuit controlling respiration in vertebrates depends on such pacemaker neurons.

Other rhythms result from synaptic interactions among neurons that are not themselves rhythmically active. In this case, the rhythms are considered to be an emergent property of the circuit, resulting from interactions among the neurons of the circuit and the properties of individual neurons. In all cases, the specific timing of the circuit elements results from the synaptic

Central pattern generators

How the nervous system achieves alternate, correctly timed contraction of extensor and flexor muscles.

connectivity among the circuit neurons.

Does the existence of central oscillatory circuits mean that reflexes and sensory feedback are unimportant for movement? Emphatically not. Sensory input is clearly important in shaping rhythmically generated movements so that they are appropriate to what the animal is doing at the time. Walking and swimming call into play neurons found in many brain regions, as well as the spinal-cord circuits and peripheral sensory neurons. It is the interplay among all these different structures that allows both relatively simple movements, such as walking over level terrain, as well as complex sequences of movements, such as in dancing or mountain climbing. All central pattern-generating circuits are subject to modulation by neurotransmitters and hormones, which can alter the frequency and intensity of the motor patterns produced by acting on specific subsets of circuit neurons.

New therapies to improve recovery after damage to the spinal cord could depend crucially on an appreciation that the spinal cord contains circuitry important for the generation of rhythmic movement. In principle, some rhythmic movement could be re-established below sites of spinal-cord lesion if the appropriate neuromodulatory environment were supplied to the central pattern-generating circuitry in the spinal cord. There are hopes that treadmill training will help to maintain the integrity of the neural circuits below lesions and facilitate functional recovery.

Rhythms are found in many areas of the brain, although their roles are not always clear. This makes it all the more important to understand the neuronal mechanisms by which rhythms with clear behavioural meaning are generated. ■

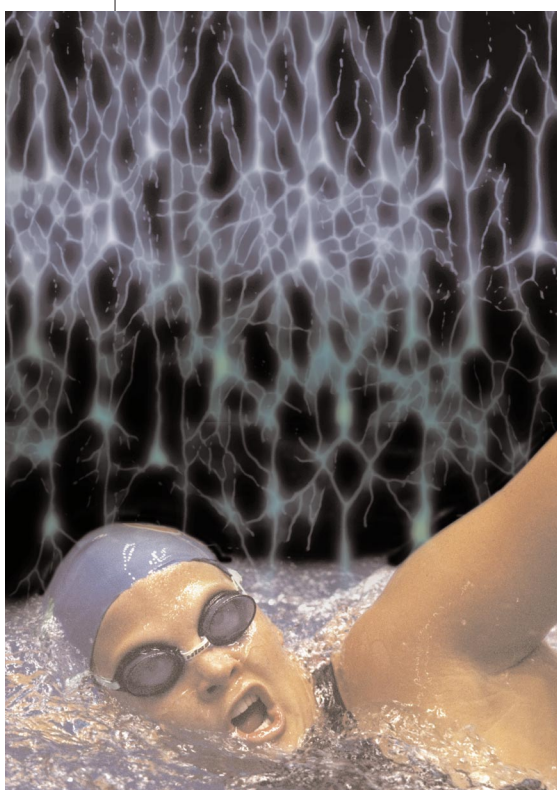
Eve Marder is at the Volen Center and Biology Department, Brandeis University, Waltham, Massachusetts 02454-9110, USA.

FURTHER READING

Marder, E. & Calabrese, R. L. Principles of rhythmic motor pattern generation. *Physiol. Rev.* **76**, 687–717 (1996).

Stein, P. S. G., Grillner, S., Selverston, A. I. & Stuart, D. G. (eds) *Neurons, Networks, and Motor Behavior* (MIT Press, Cambridge, MA, 1997).

Delcomyn, F. *Foundations of Neurobiology* (W. H. Freeman, New York, 1998).



Pop-up disaster

Wayne Thatcher

New analyses of old data show that a devastating earthquake of 1897 stemmed from slippage of a hitherto unknown fault. That event illustrates what might happen elsewhere today.

During the past decade, satellite technology has spawned a revolution in measuring the surface deformation caused by earthquakes. Yet as Bilham and England¹ show on page 806 of this issue, classical ground-based surveying methods in use as long as 150 years ago can still yield astounding results.

The Global Positioning System (GPS) constellation of satellites allows repeated measurements of ground-survey markers that are precise to a few millimetres. Complementing GPS, interferometric synthetic aperture radar (InSAR) imaging uses repeat-pass data from radar satellites to obtain complete ground-displacement maps of 100 km by 100 km regions with 100-metre spatial resolution that are accurate to 10 millimetres or better. I use both of these methods and understand their principles. Yet it still seems absurd that a satellite 800 km or more above Earth's surface, travelling at 7 km per second, should be able to provide such precise measurements of surface displacements. In contrast, ground-based triangulation, involving repeated angle measurements made with a theodolite, can typically resolve relative motions of only hundreds of millimetres over tens of kilometres.

How then can the humble surveyor's theodolite be mightier than the latest satellite technology? The answer, of course, is that, until time travel is perfected, the archive of

theodolite-derived historical data is unique. It can be mined to study the rare large earthquakes that have by chance been recorded using methods that are now largely extinct.

Using triangulation surveys carried out between 1862 and 1936, Bilham and England show that the great Assam earthquake of 12 June 1897 occurred on a previously unknown fault, which they christen the Oldham fault. The earthquake was of magnitude 8.1. The fault is located near the northern edge of the Shillong plateau in northeastern India (see map on page 806), and did not rupture the Earth's surface. Its existence can, in retrospect, be seen as necessary to explain the enigmatic existence of the plateau, an isolated, uplifted block of crust lying 200 km south of the Himalayas. This block has a steep southern flank bounded by the Dauki fault, which may be able to sustain an earthquake as large as the 1897 event, thus threatening the city of Dhaka, Bangladesh, with a population of over three million, that lies 150 km to the south of the plateau.

The 1897 earthquake holds a special place in the history of seismology. R. D. Oldham, an Englishman who led the Geological Survey of India, studied its effects and published an extensive, well-illustrated monograph² of his observations that has been much cited by succeeding generations of seismologists³. The earthquake was particularly destructive,



Figure 1 Aftermath: devastation in Muktagachha, photographed in 1897 by G. E. Grimes.

and devastated a huge region of northeastern India (Fig. 1). Observations that heavy objects such as large boulders had been thrown clear of the ground were the first documented evidence that earthquake ground motions can exceed the force of gravity. Contemporary reports in *Nature* are summarized in Box 1.

What was unusual about this earthquake was its location. Most earthquakes of this magnitude occur at the boundaries of the 15 or so large plates, some 100 km thick, that comprise the Earth's surface. Although frictional forces lock their boundaries most of the time, continued relative motion of the plates builds up forces that eventually exceed

Box 1 *Nature* reports the 1897 earthquake

An early description of the earthquake appeared in *Nature* on 17 June 1897. A week later came the report reproduced here, and a note from Thomas Heath of the Royal Observatory, Edinburgh, on violent oscillations of a pendulum at the observatory caused by the earthquake. Further evidence from pendulum records at the Royal Geodetic Observatory of Rocca di Papa, Rome, described by Professor P. Blaserna, was reported second-hand on 12 August. Finally for that year, an item published on 16 September pointed out that collation of findings about the earthquake by R. D. Oldham was underway, and that: "The earlier reports as to the damage were exaggerated; but it was nevertheless very serious, as can be readily believed, as the cylinder seismometer at Shillong indicates that the shock consisted of an oscillation of 7.4 inches at the rate of 60 times a minute."

"FURTHER reports of the earthquake in India, on June 12, show that the disturbances were felt over a very extensive area. Considerable damage was done to public and private buildings at Calcutta, and many places in the provinces of Bengal and Assam have suffered very severely. The complete breakdown of telegraphs and other means of communication delayed the reports from Assam regarding the effects of the earthquake in that province. They are now, however, coming in, and present a terrible state of affairs. Reuter reports that at Shillong the shocks of earthquake were so severe and prolonged that everything was levelled with the ground, and many people lost their lives. Gauhati is in ruins. The roads are broken up into chasms, and the railway has disappeared, but no lives were lost. At Goalpara, on the Brahmaputra, the earthquake produced a wave of water which destroyed the bazaar and all pakka buildings. The country is covered with fissures, from which mud and sand are constantly spurting. At Dhubri all pakka buildings have been demolished. The river bank has subsided, the country is flooded, and the crops are ruined. Both at Goalpara and Dhubri there has been serious loss of life. Several places are cut off from communication, and nothing has been heard from them. Throughout almost the whole of the province of Assam bridges have been destroyed and the roads rent with fissures, so that communication by road is impossible, and the telegraph cannot be used. Further information from Assam shows that the earthquake was slight in Silchar, Kohima, Manipur, Dibrugarh, and Sibsagar; violent at Barpeta, Nowgong, and Mangaldai, in addition to the places already reported. The Chief Commissioner estimates the mortality from the earthquake in the Cherra Hills at between 4000 and 6000, but details have not yet been received."

STEINBRIDGE COLLECTION, EERC, UNIV. CALIFORNIA, BERKELEY

the frictional strength of the boundary faults, producing sudden fault slips that manifest themselves in the form of roughly periodic earthquakes.

One such plate boundary occurs 200 km north of the 1897 event, where the Indian plate dives down beneath the Eurasian plate at the Himalayan mountain front. Bilham and England explain the Shillong plateau as an effect of two unusual local circumstances. The weight of the Himalayas to the north and the thick sediments of the Bengal fan to the south create a bow-shaped flexural bend of the Indian plate in Assam. These bending stresses, superposed on the compressed Indian plate, are relieved by the Shillong plateau 'popping up' by earthquake slip on its bounding faults (see Fig. 3 on page 808).

Events such as that of 1897 are comparatively rare because the earthquake slip is large and long-term fault slip rates are low. With an estimated slip of 18 m and a guesstimated fault slip rate of 2–4 mm per year, events like that the Assam earthquake should recur only every 4,000–8,000 years or so, according to Bilham and England's calculations. At plate boundaries, slip rates are an order of magnitude greater and earthquake slip is generally less, so great earthquakes occur on each fault segment every few hundred years.

Despite their idiosyncratic causes and relative infrequency, earthquakes such as the Assam event are receiving increasing attention, both because of their intrinsic interest and because of the hazards they pose to densely populated regions. Although these events are not quite plate-boundary earthquakes, they ultimately owe their existence to the dynamism of the boundary region. Plate-boundary processes generate stresses and create and destroy topography, producing local forces that must be balanced by deformation and earthquakes. Three examples illustrate the diversity of these processes.

The Seattle fault⁴ is one of a series of structures oriented east–west lying landward of the north–south-oriented Cascadia subduction zone, the boundary between the Juan de Fuca and North American plates in the northwestern United States. The east–west-oriented faults result from compression arising from northward movement of the small Oregon Coast microplate impinging on the rigid buttress of the British Columbia Coast Ranges⁵. The Seattle fault last slipped in about AD 900, and it ruptures much less frequently and causes smaller earthquakes than does the Cascadia subduction boundary. But as its name implies, this fault cuts directly through downtown Seattle, and a repeat of the AD 900 event poses a greater hazard than a larger Cascadia event. (The magnitude 6.9 earthquake of 28 February 2001 was a different type of event again — it occurred at a depth of 52 km in the subducting Juan de Fuca plate that dives beneath Seattle.)

Much of California's Los Angeles basin is

underlain by faults that, like the Oldham fault, do not yet penetrate to the Earth's surface⁶. They result from compression due to an easterly bend of the San Andreas fault, which takes up most of the relative motion between the Pacific and North American plates. Again, the proximity of these buried faults to metropolitan Los Angeles makes them arguably more hazardous than California's proverbial 'big ones' on the San Andreas.

Finally, the Bhuj earthquake, which occurred in India on 26 January 2001, underlines the importance of better understanding the causes of potentially destructive earthquakes happening near plate boundaries. This event, of magnitude 7.8, occurred on one of a series of east–west faults which, like the

Oldham fault and those of the Los Angeles basin, did not reach the Earth's surface. The origin of this fault zone is uncertain. But its location several hundred kilometres south of the Himalayas, and 200 km north of the thick fan sediments of the Indus delta, suggests that forces similar to those invoked to explain the Shillong plateau may be involved. ■

Wayne Thatcher is at the US Geological Survey, 345 Middlefield Road, Menlo Park, California 94025, USA.

e-mail: thatcher@usgs.gov

1. Bilham, R. & England, P. *Nature* **410**, 806–809 (2001).
2. Oldham, R. D. *Mem. Geol. Soc. India* **29**, 1–379 (1899).
3. Richter, C. F. *Elementary Seismology* 49–56 (Freeman, San Francisco, 1957).
4. Bucknam, R. C. *et al. Science* **258**, 1611–1614 (1992).
5. Wells, R. E. *et al. Geology* **26**, 759–762 (1998).
6. Dolan, J. F. *et al. Science* **267**, 199–205 (1995).

Reproductive biology

Out with a bang

Stuart Humphries and David J. Stevens

Not many mammalian males die soon after mating — the phenomenon was thought to be limited to certain small marsupials. Its occurrence in a larger marsupial overturns previous ideas about why the strategy exists.

Semelparity is the so-called big-bang reproductive strategy, in which an animal invests all in a single mating season, dying soon afterwards. This is not a strategy one immediately associates with mammals. Nonetheless, semelparity does occur in terrestrial mammals, but had been thought to be restricted to small marsupials from two distantly related groups, the Didelphidae and the Dasyuridae^{1,2}. Writing in *Proceedings of the Royal Society*, however, Oakwood and colleagues³ confirm a previous report⁴ that male northern quolls (*Dasyurus hallucatus*) — relatively large marsupials that weigh more than a kilogram — all die after their first mating season. Perhaps more importantly, the authors' analysis also raises doubts about why semelparity has evolved in these groups.

Females of studied Didelphidae and Dasyuridae species have a highly synchronized, single oestrus (during which ovulation occurs) in the winter, and males commit themselves totally to obtaining mates at this time, at the expense of future mating attempts. The previous explanation for the occurrence of semelparity in these two groups rested on evidence that the males have high levels of corticosteroids (stress hormones) during the mating 'rut'. Such hormones allow the animals to break down protein stored in the body to provide energy. The argument goes that the small size of these marsupials (they weigh up to 300 g), together with winter food shortages, means that they can't lay down enough energy reserves in the form of body fat. By using



Figure 1 Dying for sex: the northern quoll. Males of this species die after a single rutting season.

body protein as an energy source instead, the males can survive long enough to mate. The physiological consequences, however, are severe, and include a general decline in body condition, as well as impaired immune and inflammatory systems. Moreover, the increased concentrations of stress hormones are usually associated with gastrointestinal ulceration. It is these consequences that ultimately kill the animals.

Oakwood *et al.*'s investigation³ of semelparity in *D. hallucatus* (Fig. 1) throws doubt on this explanation. Before breeding, males of this species have sizeable fat stores in the base of their tails. Moreover, the authors found no evidence for raised corticosteroid concentrations or gastrointestinal ulcers in these males. However, body condition was very poor, and males had the increased testosterone levels found in other semelparous marsupials. Castrated males of the semelparous marsupial mouse *Antechina*

ERIC AND DAVID HOSKING/CORBIS

stuartii survive more than one breeding season⁵. The implication is that sex hormones such as testosterone are important in controlling semelparity in marsupials. So Oakwood *et al.* suggest that protein breakdown as a result of high levels of stress hormones may be a secondary modification, rather than the primary cause of mammalian semelparity. They further propose that these marsupials are faced with as yet unexplained developmental constraints that predispose them to the evolution of semelparity.

This claim is difficult to test without a comparative analysis of different marsupial and other mammalian groups. But the confirmation of semelparity in *D. hallucatus* does raise some interesting questions. In small semelparous marsupials, but not *D. hallucatus*³, the production of sperm fails shortly before mating. For the smaller marsupials, this seems logical — as they will soon die, it makes little sense for them to continue putting valuable energy into sperm production. The fact that sperm production is still maintained during the rut in *D. hallucatus* indicates that male die-off after mating may not yet be the rule in this species. Together with the fact that female northern quolls produce fewer young after their first year, this indicates that it might be fruitful for researchers to look at the build-up of deleterious mutations that act after breeding in this species.

Perhaps more importantly, the paper³ forces us to ask why semelparity evolved in these mammals, and how we might be able to reconcile this strategy with current theories of life-history evolution. Is semelparity in *D. hallucatus* the result of a 'phylogenetic predisposition', as Oakwood *et al.* suggest? In other words, is this life-history strategy a (perhaps unwanted) legacy from an ancestral species? Or is it an adaptation to specific, and as yet undefined, environmental or ecological circumstances unique to these marsupial groups?

If semelparity is to be expected as a result of a phylogenetic predisposition, then subsequent evolution, constrained by history, may tend to be limited to making the best of a bad job. But if semelparity is an adaptation, what benefits does it provide, and how did it evolve? Sexual selection might be the driving force: females might be biased towards young, vigorous males, meaning that there would be little reproductive benefit to the survival of older males. Alternatively, circumstances might be such that the intensity of fighting between males during the rut means that only the youngest and fittest can win and survive to mate. The cost of such fighting might be reduced lifespan. These ideas, although speculative, illustrate the need for further studies to determine which ecological or evolutionary circumstances promote this extreme, and extraordinary, reproductive tactic in a relatively large mammal. ■

Stuart Humphries and David J. Stevens are in the

Division of Environmental and Evolutionary Biology, Graham Kerr Building, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK.

Stuart Humphries is also at The Freshwater Biological Association, Ferry House, Far Sawrey, Ambleside LA22 0LP, UK.

e-mail: s.humphries@bio.gla.ac.uk

Condensed-matter physics

An expanding view of plutonium

R. C. Albers

Interactions between electrons make it hard to predict the properties of exotic metals, such as plutonium. Better calculations that include a thorough treatment of electronic structure are the answer.

The equations of quantum mechanics are accurate but difficult to calculate for solids, because the detailed interactions between electrons must be taken into account. For 30 years physicists have been using an approximation known as the density functional theory (DFT), which allows them to calculate the ground-state properties of solids, such as their volume. DFT treats electrons in many-atom systems as if they behave independently, and works well for standard metals like aluminium. But it fails when the solid's electrons interact strongly. For example, metallic plutonium undergoes a huge expansion in volume when it is heated to 600 K — but DFT fails to predict this. Now, on page 793 of this issue, Savrasov *et al.*¹ correctly predict the behaviour of plutonium by using a theory that marries the DFT theory for solids with a more sophisticated treatment of electronic interactions.

Imagine taking a large number of well-separated atoms and then bringing them together to form a solid. What happens to the behaviour of the electrons? At a certain point, as the clouds of electrons surrounding the atoms begin to overlap, the outermost electrons suddenly no longer belong to a single atom (as localized electrons) but are shared by all the atoms (they become delocalized electrons). If the electron orbitals that overlap are only partly filled, the resulting solid would be a metal with electrically conducting electrons. Such delocalized electrons are itinerant — they can move freely through the solid by hopping from atom to atom. In going from isolated atoms to overlapping atoms, the material crosses a transition from being an insulator (with no itinerant electrons) to a metal. This is sometimes called a Mott transition.

In the real world it is difficult to control atoms this precisely. A good approximation can be found in the series of lanthanide elements (cerium-58 to lutetium-71) and radioactive actinide elements (thorium-90 to lawrencium-103), which include plutonium-94. Stable chemical elements have filled elec-

1. Lee, A. K. & Cockburn, A. *Evolutionary Ecology of Marsupials* (Cambridge Univ. Press, 1985).
2. Cockburn, A. in *Marsupial Biology: Recent Research, New Perspectives* (eds Saunders, N. & Hinds, L.) 163–171 (Univ. South Wales Press, Sydney, 1997).
3. Oakwood, M., Bradley, A. J. & Cockburn, A. *Proc. R. Soc. Lond. B* **268**, 407–411 (2001).
4. Dickman, C. R. & Braithwaite, R. W. *J. Mamm.* **73**, 143–147 (1992).
5. Bradley, A. J., McDonald, I. R. & Lee, A. K. *Gen. Comp. Endocrinol.* **40**, 188–200 (1980).

tron shells. Each shell has a set of electron orbitals, which can be filled with a small number of electrons. The electron shells are labelled *s*, *p*, *d* or *f*, depending on the type of electron orbitals they come from. The lanthanide and actinide elements usually have three electrons in their outer *s* and *d* shells and variable numbers of electrons in their outer *f* shell. In going up the series, from lower to higher atomic number, the number of *f* electrons increases. Do we find any interesting transitions within these series as the number of electrons changes?

The shapes of the outer three *s* and *d* electrons are broad and overlap so strongly that they are metallic-like, so these elements are metals. Conversely, *f* orbitals are closer to the atom, and so are more atomic-like (Fig. 1, overleaf). The lanthanide metal cerium, which has one *f* electron, undergoes a phase transition under pressure and temperature in which the *f* electron changes from being localized to delocalized. All the other lanthanide solids have atomic-like (localized) *f* electrons. In the actinide series the lightest solids have metallic-like *f* electrons, whereas the heavier actinides have atomic-like *f* electrons. In effect, a Mott-like transition occurs in the *f* orbitals as the atomic number changes; the outermost *s* and *d* electrons hold the material together and remain largely unchanged.

This actinide transition, in which only the *f* electrons localize or delocalize, is caused by the attractive electrostatic interaction between the *f* electrons and the positively charged atomic nucleus. As the atomic number of the nucleus increases, the attraction increases correspondingly, pulling the *f* electrons closer to the atom. This causes the shape of the *f* orbital to change from overlapping to non-overlapping (Fig. 1). The *f* orbitals are more spatially extended in the actinides than in the lanthanides, and the transition occurs midway through the series; plutonium-94 has delocalized *f* electrons and americium-95 has localized *f* electrons.

The calculations of Savrasov *et al.*¹ pro-

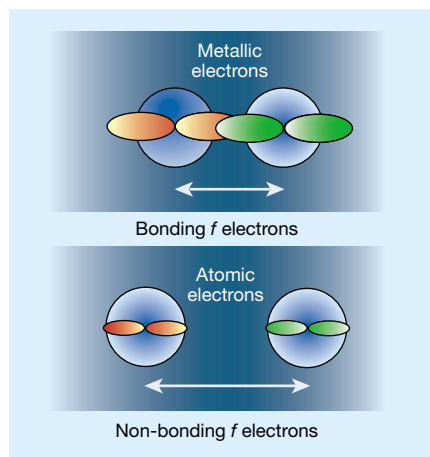


Figure 1 Metallic behaviour requires electron orbitals to be overlapping (delocalized) so that the outer electrons can move through the solid by hopping from atom to atom. Conversely, electrons in an insulator are more atomic-like (localized). Metals such as cerium and plutonium are interesting examples to study because their outer *f* electrons sit at the border between delocalized and localized behaviour, and can be driven one way or the other by changes in temperature and pressure. This figure shows the transition in the *f* orbitals, which are surrounded by a metallic sea of *s*- and *d*-orbital electrons that hold the material together.

vide a more sophisticated picture of this transition. In plutonium, temperature changes can be used to expand the atomic lattice and tune the Mott-like transition. Solid plutonium undergoes a 25% increase in volume when it changes from its room-temperature α -phase to its δ -phase at about 600 K. According to Savrasov *et al.*, the transition to the δ -phase is more complicated than previously thought. There are not just two alternatives, localized or delocalized. Instead, plutonium appears to undergo an intermediate transition that is only partly localized. This behaviour is unexpected and appears to be different from that of cerium.

The partial nature of the phase transformation shows up dramatically in the atomic volume (volume per atom) of δ -plutonium. Figure 2 shows the rough volume dependence of the series the atomic volume decreases in a parabolic manner. This effect was first explained by Friedel²: when filling electron shells in a solid, the orbitals involved in chemical bonding are filled first, causing the volume per atom to decrease. In actinides, one would expect the volume per atom to decrease until curium. Instead, between plutonium and americium the volume suddenly increases and then remains roughly constant for all higher actinides. This jump is due to the localization of the *f* electrons. For americium and higher actinides, the *f* electrons are atomic-like and do not contribute to the

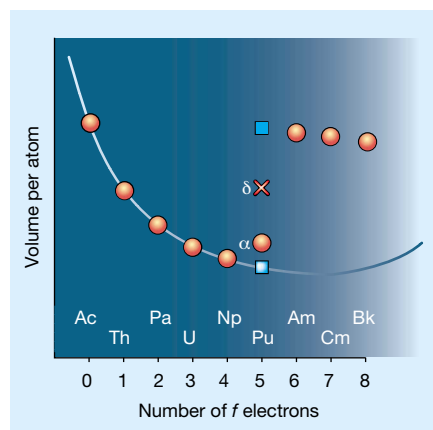


Figure 2 How the atomic volumes (volume per atom) of actinide elements, from actinium (Ac) to berkelium (Bk), change with increasing atomic number. The huge volume increase between plutonium (Pu) and americium (Am) is due to the localization of the *f* electrons because atomic-like electrons no longer contribute to chemical bonding. For americium and higher actinides the *f* electrons are localized (atomic-like), whereas the room-temperature α -phase of plutonium has delocalized (metallic-like) *f* electrons. The 25% increase in volume seen when solid plutonium is heated to 600 K (the δ -phase) has been a longstanding puzzle in condensed-matter physics. New calculations by Savrasov *et al.*¹ show that the *f* electrons in the δ -phase are half-localized, suggesting that plutonium has been arrested midway through the transition from metal to insulator. Th, thorium; Pa, protactinium; U, uranium; Np, neptunium; Cm, curium.

bonding, whereas the α -phase of plutonium has itinerant *f* electrons. The δ -phase volume is roughly halfway between α -plutonium and americium (Fig. 2), suggesting that the bonding of the *f* electrons is about half what

it should be, and that they are half-localized.

To predict this volume expansion, Savrasov *et al.* use a many-body theory called dynamical mean-field theory (DMTF) to handle the electronic interactions. Their technique is not entirely DMTF — it is built on top of a conventional density-functional theory that describes delocalized electron behaviour.

DMTF is normally used to calculate the electronic properties of materials with strongly interacting electrons, such as high-temperature superconductors and heavy-fermion materials. But Savrasov *et al.* show that such techniques can also explain the anomalous bulk behaviour of plutonium. In addition, the DMTF theory predicts an otherwise unexpected peak seen in photoemission experiments, which measure the energy distribution of the electrons. This extra peak shows the underlying atomic nature of the *f* electrons in δ -plutonium.

The technique used by Savrasov *et al.* captures qualitatively the important characteristics of δ -plutonium. Its one shortcoming is that it cannot simultaneously predict the volume expansion and the exact location of the additional photoemission peak. Further work will be needed to extend the calculations to other anomalous properties³ of plutonium caused by the partial localization of the *f* shell and to related effects in other lanthanide and actinide systems.

R. C. Albers is in the Condensed Matter Theory Group, Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. e-mail: rca@lanl.gov

1. Savrasov, S. Y., Kotliar, G. & Abrahams, E. *Nature* **410**, 793–795 (2001).
2. Friedel, J. in *The Physics of Metals* (ed. Ziman, J. M.) 361–364 (Cambridge Univ. Press, New York, 1969).
3. Cooper, N. G. (ed.) *Challenges in Plutonium Science* (Los Alamos Science, Number 26, 2000).

Biological catalysis

The hairpin's turn

Scott A. Strobel and Sean P. Ryder

A crystal structure of the hairpin ribozyme provides an atomic framework for understanding RNA catalysis, including features that may be common to both RNA and protein enzymes.

There are two types of biological catalysis: that promoted by proteins, which includes most reactions, and that carried out by RNA enzymes (ribozymes). Our understanding of the various catalytic strategies of protein enzymes has enjoyed a high-resolution structural foundation for over 30 years, but RNA enzymology has lacked anything like an equivalent footing. On page 780 of this issue¹, Rupert and Ferré-D'Amaré now provide enlightenment in the form of the high-resolution X-ray crystal structure of the hairpin ribozyme in its

catalytically active conformation. The structure provides insight into how this small RNA (comprising around 50 nucleotides) promotes cleavage of its phosphate backbone at a specific site.

The hairpin ribozyme is found in the satellite RNA of certain plant viruses. It is responsible for the autocatalytic cleavage, and subsequent rejoining in circles, of individual copies of the satellite RNA sequence during replication. At first glance its overall architecture is deceptively simple. The business end consists of two extended helices



100 YEARS AGO

The average child is capable of a quality of thinking that leads its elders, when they try to follow it, into an intellectual quagmire of inconsistency and absurdity from which they beat an inglorious retreat by angrily bidding it "not to ask silly questions". If they bid themselves not to give silly answers their request would be just... I once heard a child ask its mother, "What makes the flowers grow?" Promptly came the answer, "Jesus!" No wonder when children's intellects are muddled with such unprovable assertions that they cease to think... Granting that the results of a mother's pernicious training can be remedied in later life, it is obviously waste of valuable energy, time and money to organise an elaborate system of education to undo that which ought never to have been done. And therefore I urge that our national progress depends very largely upon "the hand that rocks the cradle": if it rocks with an intelligent purpose, it will be well with our future men; if not, then England, like Tyre, Venice and Rome... "must be led, through prouder eminences, to less pitied destruction."

From *Nature* 11 April 1901.

50 YEARS AGO

May I reply to Mr. Tyman's letter in *Nature* of February 10, p. 245? The Swedish alphabet has twenty-eight letters, of which Å is the twenty-sixth. This letter is not a modified 'A'. The relation of 'Å' to 'A' is somewhat analogous to that of 'W' to 'U'. Unfortunately the British Committee of 1937, oblivious of this, recommended the use of 'A' as an abbreviation for the ångström. The half-dozen examples quoted by Mr. Twyman of the use of 'A' in reports and books written in English since 1937 merely show that the recommendation of this Committee has quite naturally been followed. There will soon appear two English reports on symbols superseding that of 1937. One of these is sponsored by the Royal Society, the Faraday Society and the Physical Society; the other is sponsored by the British Standards Institute. Until these new reports are published, I cannot state with complete certainty what symbol they will recommend for the ångström. I can, however, affirm that both Committees charged with preparing these reports are aware of the distinction between the letters 'Å' and 'A' in Swedish.

From *Nature* 14 April 1951.

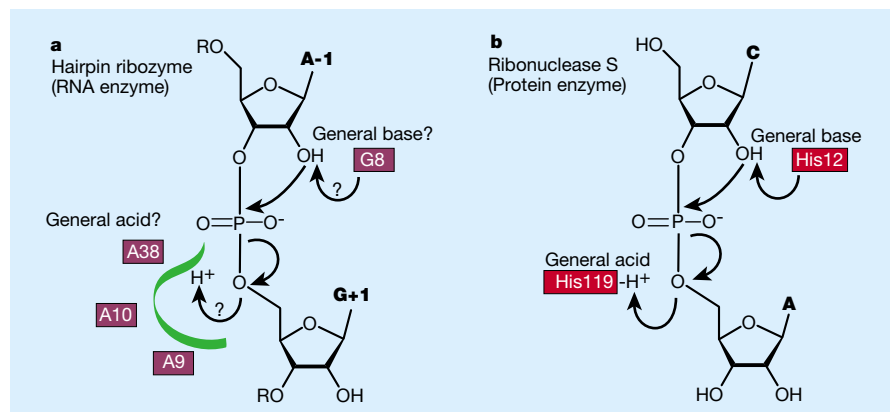


Figure 1 RNA and protein enzymes. Possible similarities between the reaction mechanisms of a, the hairpin ribozyme, and b, the protein enzyme ribonuclease S (RNase S). In the RNase S reaction mechanism, two active-site histidines (His 12 and His 119) promote general acid–base catalysis while a lysine (not shown) provides charge neutralization. Guanine and adenine bases at specific positions (G8, A9, A10 and/or A38) might have similar functions in hairpin-ribozyme catalysis. The newly solved crystal structure¹ shows that G8 contacts the 2'-OH nucleophile and the three adenines line a pocket (green) near the 5'-oxygen leaving group.

that are each interrupted by a stretch of non-complementary sequence, termed stems A and B, which end in loops (see Fig. 1b of the paper on page 781). Juxtaposition of these stems results in an RNA secondary structure that resembles a hairpin, hence the ribozyme's name.

The hairpin is small, but deciphering its catalytic mechanism has been diabolically difficult. Without insight from structural studies, it was initially assumed that the ribozyme promotes its chemical reaction using active-site metal ions by a mechanism similar to that of protein-based polymerases or phosphatases². This idea had a reassuring symmetry, in that it suggested that the RNA and protein worlds are chemically equivalent at their catalytic centres. The principle appears to be true for some larger ribozymes. But it does not apply to the hairpin and other small self-splicing ribozymes because they are fully active in the absence of metal ions^{3,4}. Although this result implied that nucleotides within the hairpin's loops are directly involved in catalysis, it was not clear how they might contribute.

Given this mechanistic uncertainty, Rupert and Ferré-D'Amaré's high-resolution crystal structure¹ comes as a boon. The hairpin ribozyme was crystallized in its docked conformation with an inhibitor at the cleavage site to prevent reaction. As seen repeatedly in other RNA structures, the loop segments are not loops at all, but are closed by a series of non-canonical interactions between the bases. These interactions create wide minor grooves that are used for tight helical packing near the active site. Both stem structures were previously solved in isolation by nuclear magnetic resonance spectroscopy^{5,6}, but their conformations are altogether different in the docked structure. For example, only two of seven non-canonically paired bases in the

isolated stem B structure are retained in the docked complex. This difference suggests that dramatic conformational rearrangements occur in both loops upon active-site assembly, a prediction supported by biochemical data⁷.

But what does the structure tell us about the catalytic mechanism of this ribozyme? As expected, the active site is devoid of metal ions and is composed entirely of RNA functional groups. Of particular importance, alignment of the two loops orients the reactive phosphate of G + 1 (guanosine + 1) for 'in line' attack by the nucleophile, the 2'-OH of A – 1 (adenosine – 1) (Fig. 1a; the numbers refer to the base distance from the reactive phosphate). This is a strong indication that the crystallized ribozyme is in its active conformation. This alignment results from the startlingly irregular geometries adopted by G + 1 and A – 1, which leave the bases unstacked from each other and rotated in opposite directions. Proper orientation of the nucleophile and leaving group may explain much of the hairpin's catalytic potential.

The structure of the hairpin ribozyme's active site does not resemble that of a protein polymerase, but is there another protein enzyme to which the hairpin can be compared? Rupert and Ferré-D'Amaré suggest that the distorted conformation around the reactive phosphate is reminiscent of ribonuclease S (RNase S), a protein that catalyses an equivalent reaction and induces similar distortions at the nucleotides surrounding the reactive phosphate⁸. However, in addition to nucleophilic alignment, RNase S uses a general acid–base mechanism of catalysis⁹. One histidine in its active site abstracts a proton from the 2'-OH nucleophile (general base catalysis), while the second donates a proton to the 5'-oxygen leaving group (general acid catalysis) (Fig. 1b).

Given the apparent conformational equivalence of the reactive phosphate in these two enzymes, does the hairpin ribozyme also use this catalytic strategy? On this point we can only speculate. RNA is inherently compromised in its ability to promote general acid–base chemistry. Unlike histidine, which has a near-neutral acidity constant (pK_a), all four of RNA's nucleotides have pK_a values that are either too high or too low to be useful in proton transfer at neutral pH. So perturbation of an active-site residue's pK_a would be required for it to function like either of the histidines in RNase S. Such perturbations are not common, but they have been observed in RNA and are proposed to play key roles in the hepatitis delta virus ribozyme and the ribosome reaction mechanisms^{10–13}. The hairpin crystal structure alone cannot define which, if any, of the bases have an unusual pK_a , but it limits the catalytically relevant candidates to four: G8, A9, A10 and A38. Of these, G8 contacts the nucleophile, whereas the three adenines line a small pocket adjacent to the 5'-OH leaving group (Fig. 1a).

If it does turn out that this small ribozyme uses a nucleotide with a perturbed pK_a for catalysis, then it may be mechanistically comparable to its much larger catalytic cousin the ribosome, which performs protein synthesis and has an active-site adenine with a

near-neutral pK_a (refs 12, 13). The molecular weight of the ribosome is more than a hundred times that of the hairpin, so this small ribozyme would define a minimal RNA element that can use this sophisticated mechanistic strategy. In this way the hairpin structure may provide clues about an ancient RNA world in which there were no protein enzymes with which to make any comparisons. ■

Scott A. Strobel and Sean P. Ryder are in the Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520-8114, USA.

e-mail: strobel@csb.yale.edu

1. Rupert, P. B. & Ferré-D'Amaré, A. R. *Nature* **410**, 780–786 (2001).
2. Steitz, T. A. & Steitz, J. A. *Proc. Natl Acad. Sci. USA* **90**, 6498–6502 (1993).
3. Nesbitt, S., Hegg, L. A. & Fedor, M. J. *Chem. Biol.* **4**, 619–630 (1997).
4. Murray, J. B., Seyhan, A. A., Walter, N. G., Burke, J. M. & Scott, W. G. *Chem. Biol.* **5**, 587–595 (1998).
5. Cai, Z. & Tinoco, I. *Biochemistry* **35**, 6026–6036 (1996).
6. Butcher, S. E., Allain, F. H. & Feigon, J. *Nature Struct. Biol.* **6**, 212–216 (1999).
7. Hampel, K. J. & Burke, J. M. *Biochemistry* (in the press).
8. Richards, F. M. *et al. Cold Spring Harb. Symp. Quant. Biol.* **36**, 35–43 (1971).
9. Deavin, A., Mathias, A. P. & Rabin, B. R. *Nature* **211**, 252–255 (1966).
10. Ferré-D'Amaré, A. R., Zhou, K. & Doudna, J. A. *Nature* **395**, 567–574 (1998).
11. Nakano, S., Chadalavada, D. M. & Bevilacqua, P. C. *Science* **287**, 1493–1497 (2000).
12. Muth, G. W., Ortoleva-Donnelly, L. & Strobel, S. A. *Science* **289**, 947–950 (2000).
13. Nissen, P., Hansen, J., Ban, N., Moore, P. & Steitz, T. *Science* **289**, 920–930 (2000).

cocaine-and-amphetamine-regulated transcript). Leptin directly suppresses the activity of the two appetite-stimulating peptides, and stimulates the activity of the appetite-reducing ones, thereby decreasing appetite. Other molecules are indirectly affected by leptin. These include melanin-concentrating hormone and a family of neuropeptides called orexins, all of which enhance appetite, as well as corticotropin-releasing hormone and oxytocin, which cause mice to eat less and to lose weight.

So, leptin is considered to be a key signal through which the hypothalamus senses the nutritional state of the body. A decrease in the amount of body fat and other energy reserves, which occurs after fasting, reduces the level of leptin. There is therefore less of this hormone around to suppress appetite-stimulating peptides and to stimulate appetite-reducing molecules, so food intake is stimulated. Higher levels of body fat and other energy reserves increase the level of leptin, leading to a reduction in food intake. This intricate mechanism is important for maintaining weight within a narrow range^{5,6}.

Di Marzo *et al.*¹ find that this weight-regulating system is even more complicated than previously thought. They show that administration of leptin reduces the levels of the endocannabinoids anandamide and 2-arachidonoyl glycerol in the hypothalamus of normal rats. Further evidence strengthens the idea that leptin downregulates endocannabinoids. In a strain of obese rats in which leptin activity is impaired, the levels of endocannabinoids are higher than normal¹. The same is true of obese *ob/ob* mice, which have an inherited lack of leptin, and of obese *db/db* mice, which have defective leptin receptors. Endocannabinoid levels are not affected in the cerebellum of these mice; this brain structure is commonly associated with motor coordination, but not with feeding. It is not yet clear how leptin might lower endocannabinoid levels specifically in the hypothalamus. There seems to be a link between leptin and the biological pathways by which the endocannabinoids are made, but further studies will be needed to pin down the details.

So far, only the food-related connection between leptin and endocannabinoids has been studied. But both types of molecule also affect other body functions, and it would be interesting to see if, and how, they interact. For example, obese *ob/ob* mice are sterile, and this is associated with low levels of leptin⁷. Indeed, these mice become fertile when treated with leptin⁷. Moreover, endocannabinoids are also involved in reproduction. Low levels of anandamide in mice are associated with the uterus being receptive to embryo implantation; higher levels coincide with the uterus being less receptive⁸. An enzyme called fatty acid amide hydrolase is needed to break down anandamide;

Physiology

A hunger for cannabinoids

Raphael Mechoulam and Ester Fride

Cannabinoids — molecules found naturally in the body, as well as in cannabis — stimulate appetite. Leptin, a hormone produced by body fat, decreases appetite. The effects of these molecules have now been linked.

The hormone leptin and neurotransmitter molecules called cannabinoids both have effects on appetite, but do they work together or separately? Writing on page 822 of this issue¹, Di Marzo and colleagues provide an answer. They show that — at least in rodents — leptin modifies the levels of cannabinoids found naturally in the body (endocannabinoids), and that endocannabinoids and leptin belong to a common system that regulates body weight.

Cannabis users are well aware of the appetite-enhancing effects of the drug. Indeed, the major active component of cannabis, a cannabinoid called Δ^9 -tetrahydrocannabinol, is used by doctors to prevent some of their patients — particularly people suffering from AIDS — from losing too much weight as a result of their illness². Fatty-acid-derived endocannabinoids (such as anandamide, isolated from pig brains in 1992; ref. 2) have many of the same properties as Δ^9 -tetrahydrocannabinol. For

example, they all bind to cannabinoid receptor proteins, found on nerve or immune cells. Anandamide also increases food intake in rats and mice^{3,4}, an effect that is prevented by a drug that selectively blocks the CB1-type cannabinoid receptor^{3,4}. The implication is that the endocannabinoids and their receptors are involved in feeding behaviour, and Di Marzo *et al.*¹ now confirm this. They find that genetically engineered 'knockout' mice that lack the CB1 cannabinoid receptor eat less than their normal littermates after being deprived of food (to enhance appetite).

The hormone leptin, meanwhile, is produced by fat tissue and affects the hypothalamus — a region in the brain that is important in weight regulation. A region of the hypothalamus called the arcuate nucleus contains neurons with receptors for two appetite-stimulating peptides (neuropeptide Y and agouti-related protein), as well as receptors for two peptides that reduce appetite (α -melanocyte-stimulating hormone and

decreased activity of this enzyme, presumably leading to increased anandamide levels, correlates with spontaneous abortion in humans⁹.

The results of Di Marzo *et al.* show that, at least with regard to appetite and feeding, the levels and effects of leptin and anandamide are inversely related. Is a similar balance important in controlling reproduction? No role for leptin after mating and fertilization (for example, in embryo implantation) has been found, and apparently it has no role in feeding during pregnancy⁷. But perhaps leptin regulates endocannabinoids in the reproductive system, as it does in nutrition.

Endocannabinoids are also involved in regulating several other processes, including stress¹⁰. Moreover, endocannabinoid-induced 'reward' effects (that is, the pleasure derived from eating tasty foods, for example) have come under scrutiny in attempts to unravel the addictive potential of cannabis¹¹. These effects are mediated by the mesolimbic brain region, and involve the neurotransmitter dopamine. Leptin also plays a role in stress¹², and may likewise affect the mesolimbic dopamine pathway and the satisfaction derived from eating¹³. Here again, as in the control of feeding, leptin and the endocannabinoids seem to work in opposite directions. Endocannabinoids activate the 'stress' (hypothalamic–pituitary–adrenal) axis¹⁰, and enhance the 'reward' experience. In contrast, leptin inhibits the hypothalamic–pituitary–adrenal axis¹² and attenuates brain reward circuitry¹³. It remains to be seen whether or not leptin affects the involvement of endocannabinoids in stress and in the reward system. But it is clear from Di Marzo *et al.*'s work¹ that much can be learned from studying the interactions between these important molecules. ■

Raphael Mechoulam is in the Department of Medicinal Chemistry and Natural Products, Medical Faculty, Hebrew University, Jerusalem 91120, Israel.

Ester Fride is in the Department of Behavioural Sciences, College of Judea and Samaria, Ariel 88437, Israel.

e-mail: mechou@cc.huji.ac.il

1. Di Marzo, V. *et al.* *Nature* **410**, 822–825 (2001).
2. Mechoulam, R., Hanus, L. & Fride, E. *Prog. Med. Chem.* **35**, 199–243 (1998).
3. Williams, C. M. & Kirkham, T. C. *Psychopharmacology* **143**, 315–317 (1999).
4. Hao, S., Avraham, Y., Mechoulam, R. & Berry, E. M. *Eur. J. Pharmacol.* **392**, 147–156 (2000).
5. Friedman, J. M. *Nature* **404**, 632–634 (2000).
6. Schwartz, M. W., Woods, S. C., Porte, D. Jr, Seely, R. J. & Baskin, D. G. *Nature* **404**, 661–671 (2000).
7. Chehab, F. F. *Trends Pharmacol. Sci.* **21**, 309–314 (2000).
8. Paria, B. C. & Dey, S. K. *Chem. Phys. Lipids* **108**, 211–220 (2000).
9. Maccarrone, M. *et al.* *Lancet* **355**, 1326–1329 (2000).
10. Zenor, B. N., Weesner, G. D. & Malven, P. V. *Life Sci.* **65**, 125–133 (1999).
11. Gardner, E. L. in *Marijuana and Medicine* (ed. Nahas, G. G. *et al.*) 187–205 (Humana, Totowa, NJ, 1999).
12. Heiman, M. L. *et al.* *Endocrinology* **138**, 3859–3863 (1997).
13. Fulton, S., Woodside, B. & Shizgal, P. *Science* **287**, 125–128 (2000).

Global change

A piece in the CO₂ jigsaw

Dorothee Bakker and Andrew Watson

A study of the year-to-year variation in net CO₂ uptake by the oceans helps in assessing the mechanisms of global climate change.

The burning of fossil fuels releases carbon dioxide into the atmosphere, where its increasing concentration promotes global warming. Some of this CO₂ is absorbed by terrestrial systems and the oceans. But the location and year-to-year variation of these 'sinks' are badly quantified. Writing in *Geophysical Research Letters*, Loukos *et al.*¹ describe an innovative approach to estimating the exchange of CO₂ between the air and the sea for the equatorial Pacific Ocean over a ten-year period. One remarkable finding, narrowing the gap between other studies^{2–6}, is that there is little variation from year to year in such CO₂ air–sea transfer in the tropical oceans.

During the period 1981–92, about 5.5 gigatonnes of carbon (Gt C) were released annually through fossil-fuel burning; 1 Gt C is 10¹⁵ g C. Of that, a variable amount, ranging from 1.4 to 5.0 Gt C, has remained in the atmosphere⁷. Research using the carbon-13 (¹³C) isotopic content of atmospheric CO₂ has ascribed this year-to-year variation to the oceanic CO₂ sink being highly variable^{2,3}. By contrast, two other studies^{4,5}, which were further constrained by measurements of the atmospheric CO₂ concentration, indicate that oceanic and terrestrial carbon reservoirs have each contributed ±0.6–1.0 Gt C to the interannual variability.

Using a different approach, Loukos *et al.*¹ have come up with a year-to-year variation of ±0.4 Gt C for tropical CO₂ air–sea transfer. The tropical oceans are an important contributor to variation in global oceanic CO₂ uptake^{4,6}, so this finding implies that oceanic involvement in interannual changes of CO₂ storage may be smaller than atmospheric studies^{4,5} suggest. But the estimate is higher than that inferred in another study using oceanic data⁶.

Conventional ocean studies estimate CO₂ air–sea exchange from the gradient in the CO₂ partial pressure (p_{CO_2}) across the sea surface and from a gas-transfer velocity, which depends on wind speed. One of the obstacles to this approach is the incomplete coverage, in both space and time, of surface-water p_{CO_2} . Instead, Loukos *et al.* take surface-water temperature and salinity as proxy measurements for dissolved inorganic carbon and alkalinity. Their rationale is that there is a close coupling between the physical and biological processes affecting these carbonate parameters in the surface waters of the equatorial Pacific Ocean.

Satellite observations and compilations of shipboard data provide good coverage of temperature and salinity. From dissolved inorganic carbon and alkalinity, Loukos *et al.* calculate the p_{CO_2} in surface water, and finally CO₂ air–sea exchange across the equatorial Pacific Ocean for the years 1982–93.

A central question in this research is whether studies based on atmospheric CO₂ data overestimate, and those based on oceanic CO₂ underestimate, the variability of the ocean CO₂ sink. One unknown factor in studies using atmospheric ¹³C values is the relative contribution of plants using specific photosynthetic systems (C₃ and C₄) to terrestrial plant growth⁸, as they affect the isotopic composition of CO₂ in different ways. Another uncertainty is the badly quantified difference in the ¹³C isotopic content of outgoing and incoming fluxes between the atmosphere and oceanic and terrestrial carbon reservoirs⁸, as fossil-fuel CO₂ emissions alter the ¹³C content of atmospheric CO₂ with time. Finally, possible errors in atmospheric transport models⁴, the relatively short time series of atmospheric ¹³C measurements, and the dominance of atmospheric monitoring stations at marine sites^{4,9} contribute further to the uncertainty of estimates based on atmospheric CO₂ data.

A major unknown in the oceanic estimates is the relationship between air–sea gas transfer and wind speed. The oceanic uptake of the radioisotope ¹⁴C, originating from natural sources and from atomic-bomb tests, provides a value for the global average transfer velocity. Otherwise there are too few direct measurements of air–sea gas exchange in the open ocean to allow the precise nature of the relationship between gas exchange and wind speed to be determined. Wind speed accounts for up to 30% of the year-to-year variation of net global oceanic CO₂ uptake⁶, so here is a considerable source of error.

Further uncertainty stems from the techniques used to obtain coverage of surface-water p_{CO_2} for an entire ocean basin. The p_{CO_2} estimates of Loukos and colleagues have a slightly lower spatial variability than corresponding shipboard observations, resulting in an underestimation of the spatial, and possibly of the temporal, variability in CO₂ air–sea exchange. Lee and colleagues' study⁶ probably underestimates year-to-year variation by normalizing surface-water p_{CO_2} data to a

single year, and by excluding the variation in biological processes, such as photosynthesis and remineralization, that affect dissolved CO_2 . So oceanic studies may indeed underestimate year-to-year variation of air-sea fluxes of CO_2 .

Most recent atmospheric and oceanic studies^{4,5,9} agree that there is a relatively large — 2 Gt C — annual uptake of CO_2 by the oceans. But there are large uncertainties in the spatial distribution of CO_2 air-sea exchange, such as the size of the CO_2 sink in the Southern Ocean^{4,9}. The location of the strong CO_2 sink in the Northern Hemisphere⁸ is of especial political and economic interest: certain atmospheric studies place it largely in terrestrial systems, notably in North America¹⁰. The lack of consensus on the distribution of CO_2 sinks, whether on land or in the oceans, let alone within specific national boundaries, poses a fundamental problem to the implementation of the Kyoto Protocol for limiting greenhouse-gas emissions. The work of Loukos *et al.*¹ contributes to this debate by suggesting that there is only moderate year-to-year change in oceanic CO_2 uptake. The results reduce the discrepancy between various estimates of the oceanic contribution to interannual variation in atmospheric CO_2 storage.

Loukos and colleagues¹ provide a good example of how improved proxy coverage of p_{CO_2} can be obtained in a specific region. However, their method of calculating dis-

solved inorganic carbon has no physical foundation and is unlikely to apply in oceanic regions outside the equatorial Pacific. In the longer term, a better quantification of the oceanic sinks for atmospheric CO_2 will require a global network of surface-water p_{CO_2} observations, combined with satellite data on temperature, algal growth and wind speed. Satellites carrying a new generation of atmospheric infrared instruments, due for launch in the next five years, should take us even closer to the near-instantaneous global detection of atmospheric CO_2 concentrations and, by inference, of the distribution of CO_2 sinks on land and in the oceans. ■

Dorothee Bakker and Andrew Watson are in the School of Environmental Sciences, University of East Anglia, Norwich NR4 7TJ, UK.

e-mails: d.bakker@uea.ac.uk

a.j.watson@uea.ac.uk

1. Loukos, H., Vivier, F., Murphy, P. P., Harrison, D. E. & Le Quéré, C. *Geophys. Res. Lett.* **27**, 1735–1738 (2000).
2. Francey, R. J., Tans, P. P., Allison, C. E., White, J. W. C. & Trolier, M. *Nature* **373**, 326–330 (1995).
3. Keeling, C. D., Whorf, T. P., Wahlen, M. & Van der Plicht, J. *Nature* **375**, 666–670 (1995).
4. Rayner, P. J., Enting, I. G., Francey, R. J. & Langenfelds, R. *Tellus B* **51**, 213–232 (1999).
5. Battle, M. *et al. Science* **287**, 2467–2470 (2000).
6. Lee, K., Wanninkhof, R., Takahashi, T., Doney, S. & Feely, R. A. *Nature* **396**, 155–159 (1998).
7. Conway, T. J. *et al. J. Geophys. Res.* **99**, 22831–22855 (1994).
8. Fung, I. Y. *et al. Global Biogeochem. Cycles* **11**, 507–533 (1997).
9. Takahashi, T. *et al. in Proc. Second Int. Symp. CO₂ in the Oceans* (ed. Nojiri, Y.) 9–15 (Centre for Global Environmental Research, Tsukuba, 1999).
10. Fan, S. *et al. Science* **282**, 442–446 (1998).

a protein called Cdc25A in a mammalian cell line in response to ionizing radiation. When active, Cdc25A removes phosphate groups from another enzyme, Cdk2, activating it and promoting progression through the cell cycle. So it was reasonable to predict that a decrease in the level of Cdc25A might lead to a failure to activate Cdk2, resulting in inhibition of the cell cycle. This hypothesis is supported by the authors' finding³ that either persistent expression of Cdc25A, or expression of a Cdc25A-resistant mutant of Cdk2, blocks the irradiation-induced S-phase checkpoint.

The next question was how does ionizing radiation enhance degradation of Cdc25A? From their previous work⁴, the authors knew that a different type of radiation, ultraviolet radiation, causes a decrease in Cdc25A levels that is dependent on the enzyme Chk1. So they now looked at whether Cdc25A could be a substrate for either Chk1 or Chk2 after ionizing radiation. The authors confirm previous observations^{5,6} that ionizing radiation activates Chk2, and find that activated Chk2 phosphorylates Cdc25A *in vitro*.

It has been suggested that mutations of Chk2 in the germ line (the tissue from which eggs or sperm are produced) are responsible for some cases of Li-Fraumeni syndrome⁷, characterized by a predisposition to cancer. Falck *et al.* show that, when they introduce either these naturally occurring Chk2 mutants, or an artificially generated, catalytically inactive mutant, into a mammalian cell line, the downregulation of Cdc25A that normally occurs in response to ionizing radiation is blocked. These mutants also inhibit the ionizing-radiation-induced S-phase delay. Conversely, introduction of normal Chk2 into a tumour cell line carrying mutated Chk2 restores ionizing-radiation-induced degradation of Cdc25A and S-phase delay. The connection between ionizing radiation, activation of Chk2 and degradation of Cdc25A was firmly established. The final piece in the jigsaw was the identification of the amino-acid site in Cdc25A that is phosphorylated by Chk2, and the demonstration of the importance of this site in the irradiation-induced phosphorylation and degradation of Cdc25A.

The authors extended their observations by linking the involvement of Chk2 in this pathway to the protein ATM (Fig. 1). Cells defective in ATM are known to lack the ionizing-radiation-induced S-phase delay. Moreover, ATM activates Chk2 during the G1 checkpoint. Falck *et al.* now show that the activation of Chk2, degradation of Cdc25A and inhibition of Cdk2 are all defective in ataxia telangiectasia cells, in which ATM is mutated (hence the name of this protein). The implication is that the phosphorylation of Chk2 by ATM is required for both G1 and S-phase delays after ionizing radiation.

Eight years ago^{8,9}, the link between the

Cell cycle

Checking two steps

Michael B. Kastan

When their DNA is damaged, cells temporarily stop multiplying to prevent the build-up of mutations. Two types of delay triggered by ionizing radiation appear to have common molecular starting points.

We thought sequencing the human genome once was daunting enough, but mammalian cells have to copy and segregate their three billion base pairs of DNA with fidelity every time they divide. If the DNA is damaged, the accuracy of this process can decrease, leading to mutations in the DNA of daughter cells. To protect themselves, eukaryotic organisms — from yeast to humans — have developed a variety of mechanisms, several of which involve transient arrests in the cell's progression through the cell-division cycle.

Although many of the mechanistic details of these 'checkpoints' have yet to be worked out, several of the proteins involved have been identified. For example, in mammalian cells, the addition of phosphate groups to the Chk2 protein (a process known as phosphorylation) by the ATM protein after exposure to ionizing irradiation contributes to

arrest in the so-called G1 phase of the cell cycle^{1,2}. Surprisingly, on page 842 of this issue³, Falck and colleagues report that the same phosphorylation event also seems to be required for an irradiation-induced delay to the S phase.

The cell cycle consists of a series of phases, starting with G1 — the resting phase that follows the previous cell division. During the next phase, S phase, the cell's chromosomes are copied. Another resting phase, G2, follows, and then eventually the cell's chromosomes are segregated and the cell divides (mitosis). In response to DNA damage, the cell can pause during progression from G1 into the S phase (at the G1 checkpoint), progression through the S phase (the S-phase checkpoint), and progression from G2 into mitosis (the G2 checkpoint).

The observation that spurred Falck *et al.*'s studies³ was a marked decrease in the level of

Daedalus

Flattening the flats

The art of making true optical flats — glass discs with surfaces polished so accurately that any unevenness is small compared with the wavelength of light — was revealed in the nineteenth century. (Until then, people had merely ground the weakest possible mirrors.) The trick is to grind three pieces of material against each other: A against B, B against C, and C against A. All three then become true flats — or at least as flat as the distortion of the Earth's gravity permits. Lens telescopes have an advantage over mirrors in that a small sag in the lens makes essentially no difference to its optical performance. Daedalus has been musing on these facts because of the high value of true flats to astronomy. The accurate distance of many stars could be deduced if a space telescope, equipped with flats, could be used as an optical interferometer.

So Daedalus now shows how make true flats. The best Earthbound flats would be made by skilled opticians, and would be flown on the space shuttle. In microgravity it should be easy to complete the ultrafine grinding of the flats against each other, giving three flats of amazing performance.

Ideally, two of these flats should be attached to a space telescope, one in its light path and the other many kilometres away. The art of stabilizing an object in space has been perfected over many decades, and Daedalus hopes that the distant flat can be turned and held so as to shine the light of a distant star unwaveringly upon its fellow flat at the space telescope. The result should be an amazingly accurate interferometric measurement of the distance of the distant star.

Daedalus has no interest in the distance of stars per se. He wants to measure the distance of stars in nearby galaxies, and thus determine the Hubble constant with accuracy. It is rather shocking that this important constant seems to vary from 60 to 85 — and that astronomers can choose whichever value they like. They can also measure it in $\text{km s}^{-1} \text{Mpc}^{-1}$, instead of s^{-1} as the rest of us would have to, but that's another matter.

An accurate Hubble constant would allow cosmologists to say clearly if the Universe were 'open' (doomed to expand for ever) or 'closed' (due to fall back on itself as the galaxies do not quite have mutual escape velocity). Daedalus likes the idea of a closed Universe, in which the galaxies only just slow before falling back. But he recognizes that a Hubble constant accurately related to the time of the Big Bang will be needed to make it work.

David Jones

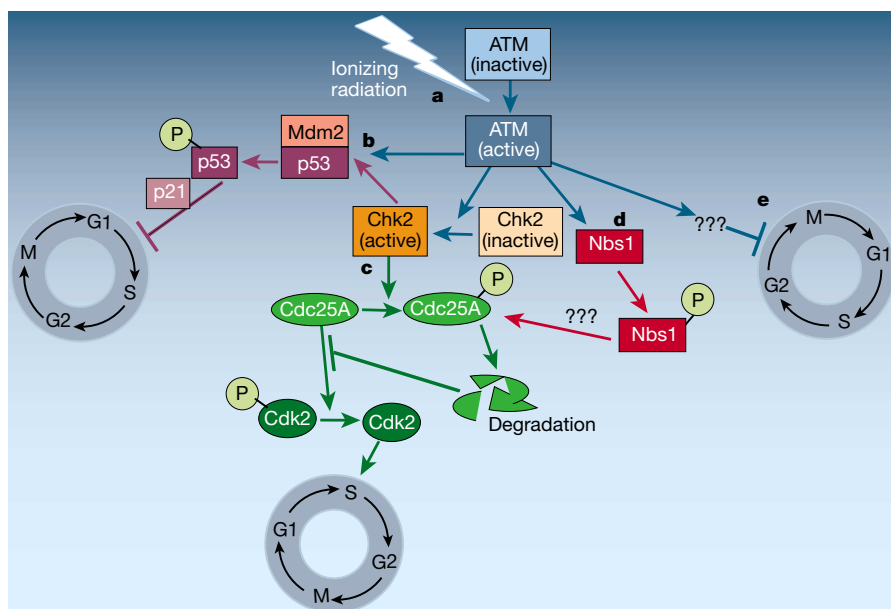


Figure 1 Steps involved in the cell-cycle checkpoints that are induced in response to ionizing radiation. The pathway investigated by Falck *et al.*³ is that shown in blue, orange and green. **a**, Ionizing radiation activates the kinase ATM, which in turn activates the kinase Chk2. **b**, This step appears to affect progression from G1 to the S phase, through phosphorylation (represented by a circled 'P') and stabilization of the p53 protein, which itself enhances the expression of the cell-cycle inhibitor p21. **c**, Activation of Chk2 by ATM also affects progression through the S phase itself, by the phosphorylation of Cdc25A. This protein is more likely to be degraded when phosphorylated. (When unphosphorylated, Cdc25A removes a phosphate group from Cdk2, enabling the initiation of DNA replication, that is, S phase.) **d**, Nbs1 is also phosphorylated by ATM and is also involved in the ionizing-radiation-induced inhibition of S-phase progression, although it is not known how it ties in to the pathway described by Falck *et al.* **e**, The targets of ATM that control progression from G2 into mitosis (M phase) have not been described.

irradiation-induced signalling pathways and the cell-cycle machinery in the G1 phase was made with the discovery that the p53 protein (a target of Chk2 in the G1-checkpoint pathway) activates the expression of an inhibitor of the cell cycle, p21. Falck *et al.*'s experiments³ linking ATM (which responds to ionizing radiation), through Chk2 and Cdc25A, to Cdk2 (which controls the cell cycle) appear to provide the same type of crucial link for S-phase progression.

But this won't be the end of the story. Other proteins will certainly be involved. For example, the protein Nbs1, which is also phosphorylated by ATM, is also required for this S-phase checkpoint^{10,11}. It will be important to work out how this particular phosphorylation event is linked to those uncovered by Falck *et al.* In the ionizing-radiation-induced G1 checkpoint, ATM phosphorylates several proteins to accomplish its goal of halting the cell cycle². Perhaps a similar scenario occurs during the control of the S-phase checkpoint (Fig. 1). It will also be important to determine whether inhibition of Cdk2 alone is sufficient to halt DNA replication, or whether other mechanisms are involved.

It is interesting that phosphorylation of Chk2 by ATM is required for both G1 and S-phase delays following ionizing radiation, even though Chk2 appears to work in differ-

ent ways in the two pathways. (In the G1 arrest, Chk2 helps to prevent the degradation of one of its targets, p53; in the S-phase delay, Chk2 enhances the degradation of another target, Cdc25A.) ATM is also required for the G2-phase delay in response to ionizing radiation. If Chk2 turns out to be involved in this delay, too, then the differences between these three checkpoints would appear to lie downstream of ATM and Chk2. Finally, Falck *et al.*'s study serves as a model for investigations of how signal-transduction pathways (in this case, one induced by DNA damage) can be linked to their functional endpoint — here, control of the cell cycle.

Michael B. Kastan is in the Department of Hematology-Oncology, St Jude Children's Hospital, Memphis, Tennessee 38105, USA.
e-mail: michael.kastan@stjude.org

- Hirao, A. *et al.* *Science* **287**, 1824–1827 (2000).
- Kastan, M. B. & Lim, D.-S. *Nature Rev. Mol. Cell Biol.* **1**, 179–186 (2000).
- Falck, J., Mailand, N., Syjuåsen, R. G., Bartek, J. & Lukas, J. *Nature* **410**, 842–847 (2001).
- Mailand, N. *et al.* *Science* **288**, 1425–1429 (2000).
- Matsuoka, S., Huang, M. & Elledge, S. J. *Science* **282**, 1893–1897 (1998).
- Brown, A. L. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 3745–3750 (1999).
- Bell, D. W. *et al.* *Science* **286**, 2528–2531 (1999).
- Harper, J. W., Adami, G. R., Wei, N., Keyomarsi, K. & Elledge, S. J. *Cell* **75**, 805–816 (1993).
- El-Deiry, W. S. *et al.* *Cell* **75**, 817–825 (1993).
- Lim, D.-S. *et al.* *Nature* **404**, 613–617 (2000).
- Zhao, S. *et al.* *Nature* **405**, 473–477 (2000).

Obituary

Claude Shannon (1916–2001)

Look at a compact disc under a microscope and you will see music represented as a sequence of pits, or in mathematical terms, as a sequence of 0s and 1s, commonly referred to as bits. The foundation of our Information Age is this transformation of speech, audio, images and video into digital content, and the man who started the digital revolution was Claude Shannon, who died on 24 February, at the age of 84, after a long struggle with Alzheimer's disease.

Shannon arrived at the revolutionary idea of digital representation by sampling a source of information at an appropriate rate, and converting the samples to a bit stream. He characterized the source by a single number, which he called entropy (a term adapted from statistical mechanics), to quantify the information content of the source. For a stream of text, Shannon viewed entropy as a statistical parameter that measured how much information is produced on average by each letter. He also created coding theory, by introducing redundancy into the digital representation to protect against corruption. If today you take a compact disc in one hand and a pair of scissors in the other, then score the disc along a radius from the centre to the edge, you will find that the disc still plays as new.

Before Shannon, it was commonly believed that the only way of achieving arbitrarily small probability of error in a communication channel, such as a telephone line, was to reduce the transmission rate to zero. All this changed in 1948 with the publication of *A Mathematical Theory of Communication*, in which Shannon characterized a channel by a single parameter, the channel capacity, and showed that it was possible to transmit encoded information at any rate below capacity with an arbitrarily small probability of error. His method of proof was to show the existence of a single good code for data transmission by averaging over all possible codes. The 1948 paper established fundamental limits on the efficiency of communication through noisy channels, but did not produce an explicit example of a good code that would achieve the highest capacity. It has taken 50 years for coding theorists to discover families of codes that come close to these fundamental limits for telephone lines.

The importance of Shannon's work was recognized immediately. According to a 1953 issue of *Fortune*, "It may be no exaggeration to say that man's progress in



Inventor, mathematician and leader of the digital revolution

peace, and security in war, depend more on fruitful applications of information theory than on physical demonstrations, either in bombs or in power plants, that Einstein's famous equation works." In fact, his work has become more important over time with the advent of deep space communication, wireless phones, high-speed data networks, the Internet, and products like compact-disc players, hard drives and modems that make essential use of coding and data compression to improve reliability and speed.

Shannon grew up in Gaylord, Michigan, and began his education at the University of Michigan, where he majored in both mathematics and electrical engineering. As a graduate student at Massachusetts Institute of Technology (MIT), his familiarity with both the mathematics of boolean algebra and the practice of circuit design produced what science historian H. H. Goldstine called "one of the most important master's theses ever written... a landmark in that it changed circuit design from an art to a science". This thesis, *A Symbolic Analysis of Relay and Switching Circuits*, written in 1936, provided mathematical techniques for building a network of switches and relays to realize a specific logical function, such as a combination lock — an electrical circuit that works if, and only if, the right combination of buttons is pressed. This work won the Alfred Noble Prize of the combined engineering societies of the USA and is fundamental to the design of digital computers and integrated circuits.

Shannon's interest in circuit design was not purely theoretical, for he also liked to

build, and his sense of play is evident in many of his creations. In the 1950s, when computers were given names like ENIAC (electronic numerical integrator and calculator), Shannon built a computer called THROBAC I (thrifty Roman-numeral backward-looking computer), which was able to add, subtract, multiply and even divide numbers up to 85 working only with Roman numerals. His study in Winchester, Massachusetts, was filled with such devices, including a maze-solving mechanical mouse and a miraculous juggling machine. Traversing the ceiling was a rotating chain, like those found at dry cleaners, from which were suspended the gowns from a score of honorary doctorates. They made a splendid sight flying around the room.

Shannon's 1941 doctoral dissertation, on the mathematical theory of genetics, is not as well known as his master's thesis, and in fact was not published until 1993, by which time most of the results had been obtained independently by others. After graduating from MIT, Shannon spent a year at the Institute for Advanced Study, and this is the period where he began to develop the theoretical framework that led to his 1948 paper on communication in the presence of noise.

He joined Bell Labs in 1941, and remained there for 15 years, after which he returned to MIT. During the Second World War, his work on encryption led to the system used by Roosevelt and Churchill for transatlantic conferences, and inspired his pioneering work on the mathematical theory of cryptography.

It was at Bell Labs that Shannon produced the series of papers that transformed the world, and that transformation continues today. In 1948, Shannon connected information theory with physics by developing a new perspective on entropy. That connection is evolving today, as others explore the implications of quantum computing, by enlarging information theory to treat the transmission and processing of quantum states.

Shannon must rank near the top of the list of major figures of twentieth century science, although his name is relatively unknown to the general public. His influence on everyday life, which is already tremendous, can only increase with the passage of time.

Robert Calderbank and Neil J. A. Sloane

Robert Calderbank and Neil J. A. Sloane are at the Information Sciences Research Lab, AT&T Shannon Research and Technology Center, Florham Park, New Jersey 07932-0971, USA. e-mails: rc@research.att.com njas@research.att.com

Herculaneum victims of Vesuvius in AD 79

The eruption's first surge instantly killed some people sheltering from the impact.

The town of Herculaneum, lying at the foot of Mount Vesuvius on a cliff overlooking the sea, was buried by a succession of pyroclastic surges and flows (currents of volcanic ash and hot gases generated by collapse of the eruptive column) during the plinian eruption of AD 79. The skeletons of 80 of 300 people who had taken refuge in 12 boat chambers along the beach have now been unearthed from the first surge deposit. We have investigated how these people were killed by this surge, despite being sheltered from direct impact, after its abrupt collapse (emplacement) at about 500 °C on the beach. The victims' postures indicate that they died instantly, suggesting that the cause of death was thermally induced fulminant shock¹ and not suffocation, which is believed to have killed many of the inhabitants of Pompeii and of Herculaneum itself.

The first surge was generated 12 hours after the eruption started². Unlike the subsequent surges, it billowed through the evacuated town of Herculaneum without damaging it and without leaving any deposit or even disturbing small and fragile heat-resistant objects. The surge advanced as a deflating current with low momentum until it reached the 20-metre cliff drop, when its basal, denser component emplaced abruptly on the beach, halted probably by hydraulic-jump effects, bursting into the waterfront chambers and enveloping the people hiding inside.

We have studied 80 intact skeletons unearthed from chambers 5 (3 skeletons out of 14), 10 (40 skeletons), 11 (5 out of 30) and 12 (32 skeletons). Their life-like stance reflects their posture at the time when the first surge emplaced. These individuals, who did not suffer mechanical impact, do not display any evidence of voluntary self-protective reaction or agony contortions, indicating that the activity of their vital organs must have stopped within a shorter time than the conscious reaction time, a state known as fulminant shock¹. The natural posture of the skeletons has been preserved by virtue of the survival of their anatomical bone connections.

The skeletons, entombed in the ash from the first and covered by the subsequent surges, are lying down or partially leaning up to a few tens of centimetres above the chamber floor, probably because of incipient 'floating'. Their positioning indicates that during emplacement the ash must have expanded slightly and then suddenly deflated, becoming denser and engulfing the bodies, cooling them and fixing them in their positions at the same time.



Figure 1 The feet of a child's skeleton recovered from a waterfront boat chamber after being entombed in the first surge of the AD 79 Vesuvius eruption. The toes show hyperflexion (flexor reflex) of the feet (chamber 12, juvenile): proximal phalanges are dorsiflexed, whereas medial and distal phalanges are plantar-flexed. The feet also show a contracture on the longitudinal axis due to both eversion and inversion, with opposition of the first and the fifth metatarsi and toes (scale rule, 10 cm).

Some of the skeletons have articulated fractures, as seen in incinerated bodies^{3,4}, and the inner skull surfaces, cranial openings and unclosed sutures are blackened from the effects of high temperature on the skull cap under increased intracranial pressure. The victims show transversal clear-cut fractures with blackened edges, and longitudinal fractures in long-bone diaphyses or flat bones, as well as cracked tooth enamel, which are also evident after incineration^{4,5}. The patterns of dental enamel cracks and of bone coloration indicate that the victims were exposed to a temperature of about 500 °C, based on observations of fire victims⁶, ancient burnt bones⁷ and human bone tissue, and on teeth heat-treated in the laboratory^{5,8}. This temperature is compatible with the estimated 480 °C determined palaeomagnetically from a tile collected from outside chamber 12.

The observed flexion of the hands and feet (caused by the thermally induced nociceptive or flexor reflex⁹) (Fig. 1) and occasional spine extension of the skeletons are evidence of instantaneous muscle contraction having occurred before the ash bed compacted. The pugilistic attitude characteristic of limb flexures that result from

tendons and muscles shortening post mortem, typical of fire victims and deaths in pyroclastic flows¹⁰, is only apparent on some of the victims.

The signs of bone carbonization and the preservation of joint connections indicate that most soft body tissues were destroyed by the intense heat and then replaced rapidly by ash. A thermodynamic calculation for a chamber filled by 30 people, on average, indicates that a sudden cooling must have occurred inside the chamber as heat from 8 cubic metres of scorching ash passed into the bodies (corresponding to 4.5 m³ of organic matter) over a contact surface area of about 20 m². The heat of the ash was just sufficient to vaporize most of the organic matter, so the initial violent vaporization caused a sudden drop in ash temperature.

This could explain why the most marked thermal effects are limited to teeth and those parts of the bones least protected by fat and tissue, and would also account for the slower disappearance of some residual soft tissue. The strongest thermal effects are exhibited by bones and teeth from people in the less crowded chamber (5), consistent with our thermodynamic results. The lack of partial thermal remanent magnetization in tiles from inside chamber 12, in contrast with the specimen outside, supports the proposed rapid drop in temperature (over a few tens of minutes¹¹).

Our findings indicate that the emplacement of the first surge caused the instant death of these 80 people as a result of fulminant shock¹. They were killed before they had time to display a defensive reaction (in less than a fraction of a second), their hands and feet underwent thermally induced contraction in about one second (estimated time based on the mean conduction velocity of nociceptive C fibres⁹), the positions of their bodies were fixed by the sudden deflation of the ash bed occurring over the next few seconds; their soft tissues were vaporized and the temperature then fell over a few tens of minutes, inhibiting the progress of the pugilistic stance and the disappearance of residual soft tissue.

Giuseppe Mastrolorenzo*, **Pier P. Petrone†**, **Mario Pagano‡**, **Alberto Incoronato§**, **Peter J. Baxter||**, **Antonio Canzanella¶**, **Luciano Fattore#**

*Osservatorio Vesuviano, Via Manzoni 249, 80123 Napoli, Italy

†Centro Musei delle Scienze Naturali, Museo di Antropologia, §Centro Interdipartimentale di Servizio di Analisi Geomineralogiche, and

#Dipartimento di Biologia Evolutiva e Comparata, Università degli Studi di Napoli Federico II,

Via Mezzocannone 8, 80134 Napoli, Italy
 ‡Soprintendenza Archeologica di Pompei,
 Scavi di Ercolano, 80056 Ercolano, Italy
 §Università degli Studi di Napoli Federico II,
 Dipartimento di Scienze della Terra,
 Largo S. Marcellino 10,
 80138 Napoli, Italy
 e-mail: incorona@unina.it
 ||University of Cambridge Clinical School,
 Addenbrooke's Hospital, Hills Road,
 Cambridge CB2 2QQ, UK

1. Brinkmann, B. et al. *Rechtsmedizin* **83**, 1–16 (1979).

2. Carey, S. & Sigurdsson, H. *Geol. Soc. Am. Bull.* **99**, 303–314 (1987).
3. Bohnert, M. et al. *Forensic Sci. Int.* **87**, 55–62 (1997).
4. Bohnert, M., Rost, T. & Pollak, S. *Forensic Sci. Int.* **95**, 11–21 (1998).
5. Yamamoto, K. et al. *Bull. Kanagawa Dent. Coll.* **18**, 55–61 (1990).
6. Holden, J. L., Phakey, P. P. & Clement, J. G. *Forensic Sci. Int.* **74**, 17–28 (1995).
7. Holck, P. thesis, Anatomical Institute, Univ. Oslo (1986).
8. Shipman, P., Foster, G. & Schoeninger, M. J. *Archaeol. Sci.* **11**, 307–325 (1984).
9. LaMotte, R. H. & Campbell, J. N. *J. Neurophysiol.* **41**, 509–528 (1978).
10. Baxter, P. J. *Bull. Volcanol.* **52**, 532–544 (1990).
11. Butler, R. F. in *Paleomagnetism* 319 (Blackwell, Boston, 1992).

High-speed swimming

Enhanced power in yellowfin tuna

Tuna are distinctive among bony fish for their elite swimming ability and their muscle anatomy, having loins of red, aerobic fibres deep within the body where other fish have only white, anaerobic fibres^{1,2}. Here we record the performance of the red muscle of yellowfin tuna *in vitro* and *in vivo* to show how this specialized muscle architecture can double the cruising power of these fish, revealing a functional link between this biomechanical design and high-speed swimming.

Active muscle must develop force and shorten in order to do work. In most fish the relative shortening, or strain, of swimming muscle is the product of body curvature and the distance of the muscle from the backbone³ (that is, strain can be accurately calculated as if the body were a homogeneous, bending beam). If a tuna's body deformed like a bending beam, then the red muscle's internal position (Fig. 1a) would limit strain and therefore the work and power produced, a seeming paradox for fish using high-performance locomotion¹.

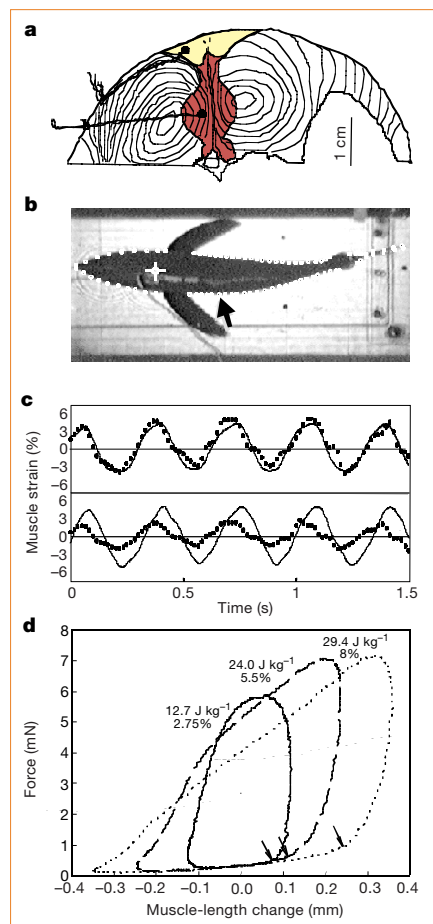
We measured muscle strain in yellowfin tuna (*Thunnus albacares*) swimming in a large water tunnel. Sonomicrometry transducers were implanted^{3,4} in superficial and in deep red muscle at the longitudinal mid-point of four animals in order to measure muscle shortening directly. We also used videography and beam theory³ to predict muscle strain by assuming that the fish bends as a homogeneous beam (Fig. 1b).

We found that for superficial red muscle there was close agreement between the measured and predicted strain amplitude and phase (Fig. 1c). However, shortening of deep red muscle measured by sonomicrometry in all fish ($\pm 5.31\%$, s.e. 0.63) was almost double that predicted by beam theory ($\pm 2.78\%$, s.e. 0.43), and it lagged behind body curvature by almost 10% of a complete cycle.

Remarkably, this phase difference produces brief intervals in which superficial muscle is lengthening while the adjacent deep red muscle is shortening, and vice

versa. There is thus a large degree of shear between superficial and deep fibres (the net strain is double that in other fish), and consequently an uncoupling of deep muscle strain from local body bending, which is not observed in other fish³.

Using work-loop techniques^{5,6} to quantify work output from muscle strips, we found that the increased strain results in significantly more work (Fig. 1d). Specifically, the average strain in deep fibres measured by sonomicrometry produces twice the work compared with the average strain predicted by local curvature (24.0 versus 12.7 joules per kg per cycle). Using even larger strains ($\pm 8\%$) does not increase work output significantly, indicating that tuna muscles are designed to work near maximum capacity at the strain they experience in the animal.



Tuna swim by restricting lateral undulations to the most caudal body segments, but maintain their sizeable red-muscle mass in the mid-body region⁷. We believe that this is possible because of the novel physical uncoupling of the action of deep red muscle from local body bending shown here. Deep red muscle at the mid-body in yellowfin tuna shortens in phase with body bending some 20% more to the posterior, supporting the idea that the tuna's complex tendon system and elongate myotomes provide a force-transmission pathway to the tail^{8,9}. We find that deep red fibres undergo strains as large as, or larger than, the strains in superficial fibres, allowing much greater work output during swimming than might be expected from their deep location.

The architecture and physiology of tuna muscle allow it to generate more work than is possible in other fish lacking this specialized anatomy. Although this increased power enables the tuna to achieve higher aerobic speeds, it must also burn more metabolic fuel. At aerobic swimming speeds, tuna do maintain a higher total metabolic rate than salmonids of similar size¹⁰, making it hard to draw conclusions about overall efficiency. But the tuna still enjoys an advantage in being able to maintain higher aerobic speeds than its prey.

Stephen L. Katz*, Douglas A. Syme†, Robert E. Shadwick‡

Figure 1 Superior performance of tuna red muscle is due to its anatomical and biomechanical design, as well as to its physiology. **a**, Tuna axial muscle in cross-section; sonomicrometer crystals are shown as black dots with trailing wires. In tuna and non-tuna fish, red muscle forms a wedge close to the skin (yellow), but tuna have red muscle deep in the myotome as well (red). **b**, Video image of a yellowfin tuna swimming at 2.3 body lengths per second. Overlaid digitized points (magnified $\times 5$) were used to calculate body margins and the curvature of the body midline³ and to predict muscle strain from beam theory. Cross, dorsal reflective position marker. Arrow, mid-body position of crystals. **c**, Comparison of red-muscle strain over 4 tailbeats as predicted from videography by beam theory³ (dots) and measured from sonomicrometry (full lines) for tuna swimming at 2.9 body lengths per second (statistics calculated for a minimum of eight tailbeats). Top, peak strain in superficial muscle: predicted, $\pm 4.79\%$; measured, $\pm 4.47\%$ (predicted strain lags behind measured by only 4.69 deg of phase). Bottom, peak strain in deep red muscle: predicted, $\pm 2.41\%$; measured, $\pm 5.46\%$ (measured strain lags behind predicted by 30.7 deg). Orientation of sonomicrometry transducers was verified post mortem. **d**, Work loops from segments of deep red muscle. The muscle was stimulated phasically during sinusoidal length oscillations; onset (arrows) and duration of stimulation were adjusted to give maximum net work (loop area), and were similar to activation timing *in vivo*⁹. Loops run anticlockwise, so that the force is relatively low during lengthening and high during shortening. Frequency of the length oscillation (tailbeat frequency) was 3 Hz. Peak amplitude of imposed length change (muscle strain as a percentage of resting length) was $\pm 2.75\%$ (solid line), $\pm 5.5\%$ (broken line), or $\pm 8\%$ (dashed line). The net work done per cycle is shown with the strain for each loop. Muscle resting length, 4.4 mm, corresponding to zero length change on the graph.

*Biological Laboratory for Open Water Mechanical Engineering, 3241 Ocean Drive, Oxnard, California 93035, USA

e-mail: steve.katz@noaa.gov

†Department of Biological Sciences, University of Calgary, Calgary AB T2N 1N4, Canada

‡Marine Biology Research Division, Scripps Institution Of Oceanography, La Jolla, California 92039-0202, USA

1. Fierstine, H. L. & Walters, V. *Mem. S. Calif. Acad. Sci.* **6**, 1–31 (1968).
2. Magnuson, J. J. in *Fish Physiology* Vol. VII (eds Hoar, W. S. & Randall, D. J.) 239–313 (Academic, New York, 1978).
3. Katz, S. L. *et al. J. Exp. Biol.* **202**, 529–541 (1999).
4. Shadwick, R. E. *et al. J. Exp. Biol.* **202**, 2139–2150 (1999).
5. Josephson, R. K. *J. Exp. Biol.* **114**, 493–512 (1985).
6. Altringham, J. D. & Johnston, I. A. *J. Exp. Biol.* **148**, 395–402 (1990).
7. Graham, J. B. *et al. Can. J. Zool.* **61**, 2087–2096 (1983).
8. Westneat, M., Hoese, W., Pell, C. A. & Wainwright, S. A. *J. Morphol.* **217**, 183–204 (1993).
9. Knowler, T., Shadwick, R. E., Katz, S. L., Graham, J. B. & Wardle, C. S. *J. Exp. Biol.* **202**, 2127–2138 (1999).
10. Korsmeyer, K. E., Dewar, H., Lai, N. C. & Graham, J. B. *Comp. Biochem. Physiol. B* **113**, 45–56 (1996).

Environment

Superselective clay for radium uptake

A legacy of the nuclear industry that threatens many areas of the world is contamination by radioactive radium from the solid and liquid wastes left after extracting uranium from its ores. Here we describe a highly charged, synthetic clay that specifically removes radium and immobilizes it at room temperature for safe disposal. This material may prove to be useful in decontaminating drinking water wells and soils of this toxic isotope.

Synthetic cation exchangers have been tailored for decontaminating the environment of radioactive isotopes such as ^{137}Cs and ^{90}Sr (refs 1–3) and immobilizing them through electrostatic bonding^{2,3} at room temperature. There is also a need for selective exchangers for the uptake of radium because not only is this isotope extracted into groundwater from uranium wastes but naturally occurring uranium may also contaminate drinking-water wells. Radium must be filtered out of drinking water to protect the populace from cancer and other health risks^{4,5}.

Two synthetic micas of nominal composition $\text{Na}_4\text{Al}_4\text{Si}_4\text{Mg}_6\text{O}_{20}\text{F}_4 \cdot x\text{H}_2\text{O}$ ('Na-4-mica') and another two of nominal composition $\text{Na}_2\text{Al}_2\text{Si}_6\text{Mg}_6\text{O}_{20}\text{F}_4 \cdot x\text{H}_2\text{O}$ ('Na-2-mica') were prepared from sol-gel precursors³ (gel method) or by using metakaolinite and MgO as precursors⁶ (kaolinite method) in sodium fluoride at 890 °C. We compared the uptake properties of several natural montmorillonite clays and of these synthetic clays.

Selective uptake of radium was determined using 0.5 M, 1.0 M or 2.0 M NaCl

Table 1 Radium uptake by different clays

	0.5 M NaCl	1 M NaCl	2 M NaCl
Na-4-mica*	1,034 ± 64	234 ± 15	58 ± 3
Na-4-mica†	856 ± 80	207 ± 7	55 ± 2
Na-2-mica†	787 ± 42	167 ± 4	45 ± 2
Na-2-mica*	544 ± 9	113 ± 4	30 ± 0.5
Montmorillonite, Wyoming	12.4 ± 6.1	4.8 ± 0.3	4.8 ± 0.3
Montmorillonite, Arizona	11.5 ± 4.0	9.6 ± 0.6	7.2 ± 0.1
Montmorillonite, Kunipia	44.4 ± 0.2	16.9 ± 0.3	6.8 ± 0.2

Uptake of radium by different clays from solutions containing high concentrations of NaCl is measured as the distribution coefficient K_d , in ml g^{-1} , defined as the ratio of the amount of radium sorbed per gram of solid to the amount of radium remaining per millilitre of the solution^{1–3}.

*Synthesized by the kaolinite method; †synthesized by the gel method.

containing $^{226}\text{RaCl}_2$ at about 6×10^{-9} M. These solutions have extremely large equivalent ratios of $[\text{Na}^+]/[\text{Ra}^{2+}]$ (0.6×10^8 , 1.2×10^8 and 2.4×10^8 in 0.5, 1.0 and 2.0 M NaCl solutions, respectively). Aliquots of the solution (20 ml) and sorbent powder (20 mg) in a centrifuge tube were equilibrated at room temperature. After 24 h, the solid and solution phases were separated and the ^{226}Ra concentration in solution was determined. The selective uptake of radium is expressed as the distribution coefficient K_d .

Table 1 shows that Na-4-mica is the best of all the clays tested at taking up radium from solutions containing 60 to 240 million times more sodium than radium. This mica removed $95.4 \pm 0.3\%$ radium from 0.5 M NaCl solution. Its high charge density and special structure of offset layer stacking means that Na-4-mica prefers to take up fewer hydrated monovalent and divalent ions³, which accounts for its superselectivity towards radium.

To investigate this selective uptake by Na-4-mica, we determined the equilibrium Na/Ba exchange isotherm using barium as a substitute for radium⁷. The shape of this isotherm^{2,3} also reveals the nature of the uptake process⁸. Results shown in Fig. 1 indicate that the mica prefers Ba^{2+} in the region of low concentration (initially steep curve), but then exchange stops (flat curve caused by the collapse of the interlayer space)^{3,8}. This type of interlayer collapse occurs in layered clays because of electrostatic bonding^{2,3} with ions of low hydration such as Cs^+ , Rb^+ and Sr^{2+} .

Calculations from the equilibrium isotherm (Fig. 1) show that about 25% of the available barium enters the Na-4-mica interlayers, corresponding to a capacity of

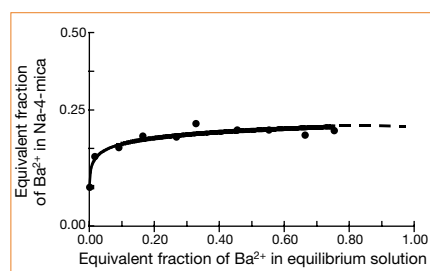


Figure 1 Selective ion-exchange isotherm for Na-4-mica in mixed NaCl/BaCl₂ solutions (barium used as a simulant for radium).

115 mequivalents per 100 g, before collapse of the interlayers prevents further exchange. Radium ion, which is less hydrated than barium ion, should not be taken up to the same extent as barium as it is more likely to be trapped (immobilized) in the interlayers after fractional exchange owing to electrostatic and steric factors.

This special superselective mica could be used to help solve not only the radium-separation problem, but also its fixation and safe disposal.

Sridhar Komarneni*, **Naofumi Kozai†**, **William J. Paulus‡**

*Department of Agronomy and Materials Research Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802, USA
e-mail: komarneni@psu.edu

†Department of Environmental Science, Japan Atomic Energy Research Institute, Tokai, Ibaraki 319-1195, Japan

‡AC Rochester Division of General Motors Corporation, Department 32-84, 1300 North Dort Highway, Flint, Michigan 48556, USA

1. Komarneni, S. & Roy, R. *Nature* **299**, 707–708 (1982).
2. Komarneni, S. & Roy, R. *Science* **239**, 1286–1288 (1988).
3. Paulus, W. J. *et al. Nature* **357**, 571–573 (1992).
4. Finkelstein, M. M. *Can. Med. Assoc. J.* **151**, 565–571 (1994).
5. Lyman, G. H., Lyman, C. G. & Johnson, W. J. *Am. Med. Assoc.* **254**, 621–626 (1985).
6. Komarneni, S., Pidugu, R. & Amonette, J. E. *J. Mater. Chem.* **8**, 205–208 (1998).
7. Jurado Vargas, M., Olguin, M. T., Ordóñez Regil, E. & Jiménez Reyes, M. J. *Radioanal. Nucl. Chem.* **218**, 153–156 (1997).
8. Breck, D. W. *Zeolite Molecular Sieves* (Krieger, Malabar, 1984).

Neanderthal DNA

Not just old but old and cold?

The successful retrieval of ancient DNA from two geographically dispersed Neanderthal skeletons^{1,2} has fuelled a demand for more Neanderthal DNA sequences for analysis. However, these exceptionally well-preserved specimens were geologically young and the mean annual temperature of their cave sites low, so the survival of this ancient DNA could have been due to unusually favourable conditions. Here we calculate the thermal history of a range of Holocene and Pleistocene bones whose DNA quality has been tested and find that in only very few sites with Neanderthal remains is the preservation of

DNA likely to match the quality of that from the skeleton found at Mezmaiskaya Cave². We recommend that any additional Neanderthal destined for destructive analysis should be carefully selected, taking into account its integrated thermal history.

We calculated the likely deterioration of DNA for a series of Holocene fossils and at Pleistocene sites^{1–8} for which we had reliable regional palaeoclimatic reconstructions^{9–11}. We assumed that depurination is the principal pathway of DNA decomposition in most environments¹² (activation energy, $E_a = 127 \text{ kJ mol}^{-1}$) and have taken account of altitude and the integrated regional palaeotemperature data in the calculations. For each sample, the expected deterioration in DNA by depurination alone is expressed in terms of its 'thermal age', which is the number of years required, at a constant 10 °C, to produce the degradation calculated from its thermal history.

Figure 1 shows that bones with thermal ages greater than that of the Feldhofer Neanderthal¹ (17,000 years at 10 °C) failed to yield DNA. The material from Mezmaiskaya², although this is at a lower latitude than Feldhofer¹, was cooler because of its altitude (1.3 km; ref. 2) and so has a lower thermal age. High altitude is a common factor in low-latitude sites that have yielded ancient DNA^{6,7} and is anticipated from thermal age analysis. The failure to amplify DNA samples in the polymerase chain reaction from specimens from two sites whose thermal ages are both younger than 17,000 years at 10 °C demonstrates that thermal age is not the only factor that preserves DNA quality for successful amplification. Nevertheless, our analysis indicates that the original Feldhofer Neanderthal DNA sequence¹ represents the current technical limit for retrieval.

When we subjected other Neanderthal sites from northwestern Europe to the same analysis, only nine (of 39) cave sites are thermally younger than Feldhofer, and none strikingly so (see table in supplementary information). Despite the importance of gaining further insight into Neanderthal genetics, our results reveal that careful consideration is necessary before embarking on the destructive analysis of remains from all but a handful of sites.

Colin I. Smith*, Andrew T. Chamberlain†, Michael S. Riley‡, Alan Cooper§, Chris B. Stringer||, Matthew J. Collins*

*Fossil Fuels and Environmental Geochemistry (Postgraduate Institute), NRG, Drummond Building, University of Newcastle, Newcastle upon Tyne NE1 7RU, UK
e-mail m.collins@ncl.ac.uk

†Department of Archaeology, University of Sheffield, Northgate House, Sheffield S1 4ET, UK

‡School of Earth Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

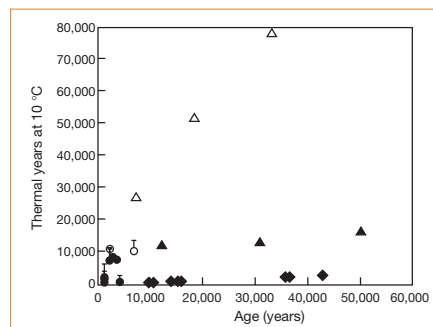


Figure 1 The success of amplifying ancient DNA by polymerase chain reaction^{1–8} is related to its thermal age — the thermal age of the original Neanderthal DNA amplification¹ represents the limit using current techniques. Filled symbols represent sites where ancient DNA has survived, clear symbols sites where it has not: diamonds, permafrost sites; triangles, cave sites; and circles, open sites. Thermal ages of open sites are calculated using mean annual air temperatures, with the upper range representing the effect of temperature fluctuation (estimated from monthly averages). The relation between air temperature and permafrost surface temperature is difficult to predict¹³, but these problems are largely offset because the rate of depurination is apparently weakly temperature-sensitive in frozen samples¹⁴. Methodological details are available from the authors.

§Departments of Biological Anthropology and Zoology, University of Oxford, Oxford OX2 6QS, UK
||Department of Palaeontology, The Natural History Museum, London SW7 5BD, UK

1. Krings, M. *et al.* *Cell* **90**, 19–30 (1997).
2. Ovchinnikov, I. *et al.* *Nature* **404**, 490–493 (2000).
3. Cooper, A. *et al.* *Science* **277**, 5329–5332 (1997).
4. Leonard, J. A., Wayne, R. K. & Cooper, A. *Proc. Natl Acad. Sci. USA* **97**, 1651–1654 (2000).
5. Höss, M., Pääbo, S. & Vereshchagin, N. K. *Nature* **370**, 333 (1994).
6. Cooper, A. *et al.* *Nature* **381**, 484 (1996).
7. Fleischer, R. C., Olson, S. L., James, H. F. & Cooper, A. C. *Auk* **117**, 1055–1060 (2000).
8. Colson, I. B., Bailey, J. F., Vercauteren, M. & Sykes, B. C. *Ancient Biomol.* **1**, 109–117 (1997).
9. Vandenbergh, J., Coope, R. & Kasse, K. *J. Quat. Sci.* **13**, 361–366 (1998).
10. Guiot, J. *et al.* *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **103**, 73–93 (1993).
11. Elias, S. A. *Quat. Res.* **53**, 229–235 (2000).
12. Lindahl, T. & Nyberg, B. *Biochemistry* **11**, 3610–3618 (1972).
13. Zhang, T., Osterkamp, T. E. & Stamnes, K. *Permafrost Periglacial Process.* **8**, 45–67 (1997).
14. Osborne, M. R. & Phillips, D. H. *Chem. Res. Toxicol.* **13**, 257–261 (2000).

Supplementary information is available on Nature's website at <http://www.nature.com> or as paper copy from the London editorial office of Nature.

Ovchinnikov *et al.* reply — Smith *et al.* have shown an interesting correlation between the thermal age of Pleistocene–Holocene fossils and ancient DNA retrieval and advocate using this criterion to assess the merit of subjecting ancient bones to destructive analysis, particularly any scientifically valuable bones of Neanderthals and anatomically pre-modern humans. The post-mortem DNA quality is, of course, dependent on the surrounding temperature, but — as the authors point out — many other important factors influence fossil DNA preservation, such as air and soil humidity, soil pH, phosphorus content of the soil, average tempera-

ture in different earth layers, and microbial-mediated decay, which also have to be taken into account.

The complex interaction of several factors is illustrated by differential preservation of skeletons and bones from the same site, and even between different parts of the same skeleton or mummy. In addition, some specimens recovered from permafrost or glacial conditions after present-day thawing have shown variable degrees of mitochondrial DNA preservation (for example, in the Tyrolean Ice Man¹) or they contained no recoverable mitochondrial DNA (as discovered in the Altai Princess and Warrior mummies; I.V.O. and W.G., manuscript in preparation), illustrating that low temperatures alone are not sufficient to preserve all specimens.

Relying too heavily on temperature alone could lead to important specimens being excluded from DNA analysis. The Feldhofer Neanderthal is an outlier based on this model, which would predict that the possibility of recovering fossil DNA from this specimen would be very remote. Obtaining a sample of the Feldhofer Neanderthal for destructive analysis would presumably have been even more difficult if too much significance had been placed on such a model at the time the work was carried out.

Another example for which the thermal age may not be an accurate guide to the state of preservation is the Neanderthal recovered from the Marillac Cave in France. As this has a thermal age of 30,539 years, the model would predict that there should be very little chance of fossil DNA remaining. However, the collagen yield is comparable to that of the Mezmaiskaya Neanderthal, which had very good DNA preservation^{2,3}.

In conclusion, we agree that there is a real correlation between DNA survival and thermal history, but the accuracy to which the survival of DNA can be predicted is limited. Important specimens could be overlooked if too much emphasis is placed on the thermal age alone.

Igor V. Ovchinnikov*, Anders Götherström§, Galina P. Romanovall, Vitaliy M. Kharitonov¶, Kerstin Lidén§, William Goodwin*

*Human Identification Centre, University of Glasgow, Glasgow G12 8QQ, UK
e-mail: w.goodwin@formed.gla.ac.uk

†Institute of Gerontology, Moscow 129226, Russia

‡Department of Medicine, Columbia University, New York, New York 10032, USA

§Archaeological Research Laboratory, Stockholm University, 106 91 Stockholm, Sweden

||Institute of Archaeology, Moscow 117036, Russia

¶Institute and Museum of Anthropology, Moscow State University, Moscow 103009, Russia

1. Handt, O. *et al.* *Science* **264**, 1775–1778 (1994).

2. Bocherens, H. *et al.* *J. Hum. Evol.* **20**, 481–492 (1991).

3. Ovchinnikov, I. V. *et al.* *Nature* **404**, 490–493 (2000).

The role of chaotic resonances in the Solar System

N. Murray* & M. Holman†

* Canadian Institute for Theoretical Astrophysics, 60 St George Street, University of Toronto, Toronto, Ontario M5S 3H8, Canada

† Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

Our understanding of the Solar System has been revolutionized over the past decade by the finding that the orbits of the planets are inherently chaotic. In extreme cases, chaotic motions can change the relative positions of the planets around stars, and even eject a planet from a system. Moreover, the spin axis of a planet—Earth's spin axis regulates our seasons—may evolve chaotically, with adverse effects on the climates of otherwise biologically interesting planets. Some of the recently discovered extrasolar planetary systems contain multiple planets, and it is likely that some of these are chaotic as well.

Observers have targeted solar-type stars in their searches for planets, partly motivated by the belief that these stars might harbour Earth-like planets. Although present detection techniques are unable to detect terrestrial-mass planets, they are sensitive to Jupiter-mass bodies. The original idea was that because all of these solar systems should share a common planet formation process they would result in similar collections of planets; large planets might be accompanied by smaller counterparts. However, the wide variety of orbits of the planets detected so far shows that our knowledge of planet formation is incomplete. Nevertheless, although the formation processes in other solar systems might be quite different from those that shaped our own, the long-term dynamical evolution is governed by the same principles.

These principles have been known since Newton, but their consequences for our Solar System still surprise researchers. For nearly 300 years astronomers believed that the orbits of the planets were regular and predictable. After an initial phase of formation, the number of planets, asteroids and comets was fixed. Planets neither escaped the Solar System nor collided with each other. The discovery of chaos destroyed this traditional picture.

Our Solar System provides a plethora of examples of chaotic motion. The theory of chaos has been used to explain, and in some cases predict, the location and extent of gaps in the asteroid and Kuiper belts^{1–7}. The theory predicts that irregularly shaped satellites such as Hyperion tumble chaotically⁸. The obliquity of Mars undergoes large excursions⁹; these are chaotic^{10,11}. Even the obliquity of the distant future Earth may undergo chaotic evolution in 1.5–4.5 Gyr

(refs 12–14). Furthermore, the chaos in the orbits of the giant planets, first observed by Sussman and Wisdom¹⁵, is due to a three-body resonance among Jupiter, Saturn and Uranus¹⁶; the locations of these planets cannot be predicted on a timescale longer than a few tens of millions of years. This supports numerical experiments that showed that our planetary system evolves chaotically^{15,17}. The worried reader may find some comfort in that the accompanying analytic theory predicts that no planet will be ejected before the Sun dies.

The rigid pendulum

Despite the variety and complexity of the applications mentioned above, we can introduce many of the concepts in solar system dynamics using the pendulum: phase space structure, periodic motion and stability. To describe the state of the pendulum we must specify both its position (such as the angle from the downward vertical, θ , see Fig. 1) and velocity (such as the rate of change of the angle with time, $\dot{\theta}$, or the momentum $p = m l \dot{\theta}$, where m is the effective mass of the pendulum and l is the length). The structure of the phase space is built around fixed points (see Fig. 2). When the pendulum is hanging straight down and motionless we have a fixed point. There is another fixed point when the pendulum is straight up and motionless. Readers equipped with a pencil can immediately see that these two points are different: dangling a pencil loosely

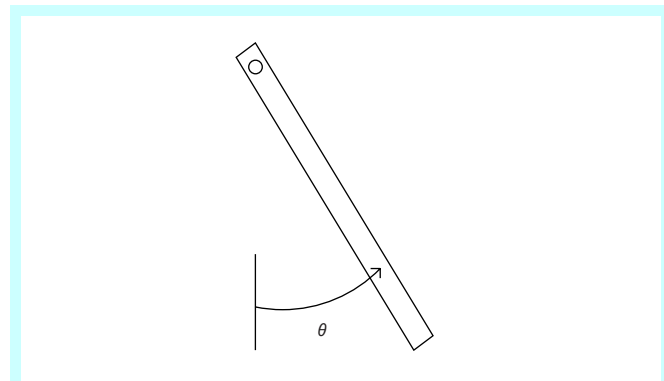


Figure 1 A rigid pendulum, with angle θ measured from the vertical. The pendulum may rotate through 360° , unlike a simple pendulum consisting of a mass suspended on a string.

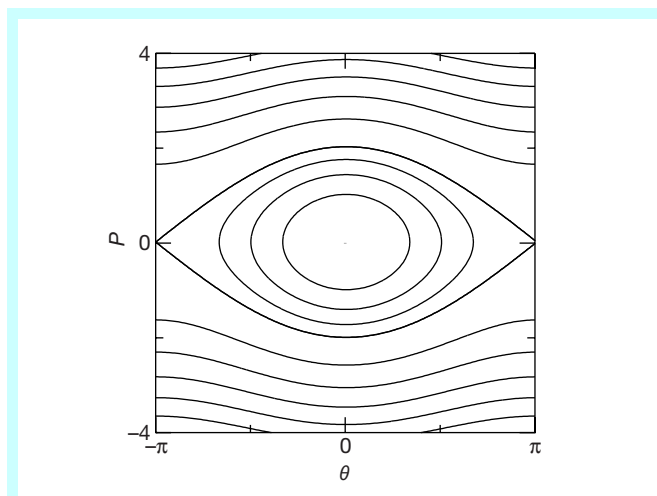


Figure 2 The motion of a rigid pendulum traces closed curves on the phase diagram, showing the angle of pendulum, θ , versus its angular momentum, $p = m l \dot{\theta}$. The stable fixed point is at $(0,0)$. The separatrix (see text) emanates from the unstable fixed point at $(\pm\pi, 0)$. The shape of the separatrix resembles a cat's eye.

between two fingers is easy, but balancing a pencil upright challenges the most dextrous.

If a pendulum at the downward fixed point is nudged it begins to oscillate or librate back and forth; this fixed point is said to be stable. If a pendulum momentarily balanced on its head is gently nudged it swings down and back up on the other side; this fixed point is unstable. If the pendulum is pushed forcefully enough it will rotate periodically. Thus, we see three main regions of the phase space: circulation in one direction, libration, and circulation in the other direction. These three regions are separated by the curve that connects the unstable fixed point to itself. This curve forms the 'cat's eye' in Fig. 2, and is called the separatrix. All trajectories inside the separatrix are said to be in resonance. Trajectories well outside the 'cat's eye' have enough kinetic energy to be unaffected by the gravitational potential. They rotate with an almost fixed angular frequency. In contrast, trajectories inside the resonance are dominated by the potential; they librate. The average value of $\dot{\theta}$ in the resonance is zero, which is a generic signature of resonant phenomena.

The separatrix is the cradle of chaos. Suppose we reach in a hand and just nudge a moving pendulum in an attempt to change the character of its motion. If the pendulum is far from the separatrix (either librating near the stable fixed point or rotating) the nudge will not change the character of the motion. However, close to the separatrix the pendulum nearly stands motionless on its head once every libration or rotation period. A nudge can then change rotation into libration, libration into rotation, or rotation in one direction into rotation in the other direction. Aside from the separatrix, all the orbits in Fig. 2 lie on closed curves. These curves are called invariant curves; similar curves appear in most dynamical systems.

In order for chaos to occur there must be at least two interacting oscillators. In the Solar System that interference is supplied by a third body, the invisible 'hand' made immanent. This interference manifests itself as zones of chaotic motion, which are the 'ghosts' of the departed separatrices. That phase space would harbour regions where trajectories are largely quasiperiodic and regions where trajectories are largely chaotic is a feature of generic dynamical systems. This characteristic is known as a divided phase space^{18,19} (see Box 1).

Kirkwood gaps

A series of remarkable features in the asteroid belt vividly illustrates

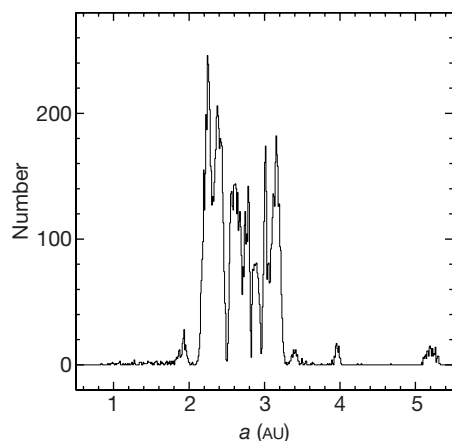


Figure 3 The histogram of asteroids as a function of semimajor axis a . We note the distinct deficit of asteroids at, for example, 2.5, 2.82, and 2.96 AU. Gaps of this type were first noted by Kirkwood in the late 1800s. They correspond to orbital resonances with Jupiter; for example, an asteroid placed at 2.5 AU will orbit with a period equal to one-third of Jupiter's (the 3:1 resonance).

the importance of dynamical chaos in the Solar System. The distribution of semimajor axes of the orbits of the asteroids contains a number of distinct gaps (see Fig. 3). These are called Kirkwood gaps²⁰. Kirkwood first identified them and noted that they occur at locations where the orbital period, T , which depends on the semimajor axis, would be of the form $(p/q)T_J$, where T_J is the orbital period of Jupiter and p and q are integers. In terms of the orbital frequencies, $n = 2\pi/T$, $qn_J - pn \approx 0$; in other words, this is a resonance. Nearly a century passed before it was explained^{2,3} how a resonance produced a shortage of asteroids, or a gap. The reason for this delay can be appreciated by making a simple estimate of the strength of a resonant perturbation. The force is proportional to the mass of Jupiter, which is $\mu_J \approx 0.001$ in units of the solar mass. Worse still, the resonant force is zero if the orbits are circular (for a planar model). If the orbits are non-circular, with an eccentricity e for the asteroid and e_J for Jupiter, the force is proportional to $e^{p-q}e_J^s$, where s is an integer between zero and $|p - q|$. For example, for a 3:1 resonance $p = 3$ and $q = 1$, three force terms are possible: e^2 , ee_J , and e_J^2 . Because $e \approx e_J \approx 0.05$, the force exerted on the asteroid by the resonance is smaller by a factor of 5×10^{-5} than the force exerted by the Sun, in the best possible case of a first-order resonance ($|p - q| = 1$).

Laplace noted even before the discovery of the Kirkwood gaps that a resonant force adds up over many orbits, but only up to half the libration period (the time to go halfway around the 'cat's eye' in Fig. 2), which is proportional to the inverse of the square root of the resonant force. The integrated force is then larger, roughly like the square root of $\mu_J e^{p-q}$. This is still a small effect, so it is difficult to see how the asteroid can be ejected from the resonance.

Even if the force added up over longer times, theorems on the behaviour of dynamical systems from the 1950s and 1960s suggested that gap formation might be difficult. For example, the motion of a typical asteroid in an imaginary solar system with very small planetary masses lies on invariant curves similar to those in Fig. 2 (see Box 1). This is a specific example of the KAM theorem²¹⁻²³. In exceptional cases, near the separatrix of a resonance, the theorem does not apply.

That the theorem is restricted to very small masses suggests a way to evade the dilemma; perhaps for realistic values, invariant curves at substantial distances from the separatrix are destroyed. Numerical integrations of a simplified model, incorporating only a single planet (Jupiter) on a fixed ellipse, showed that the invariant curves

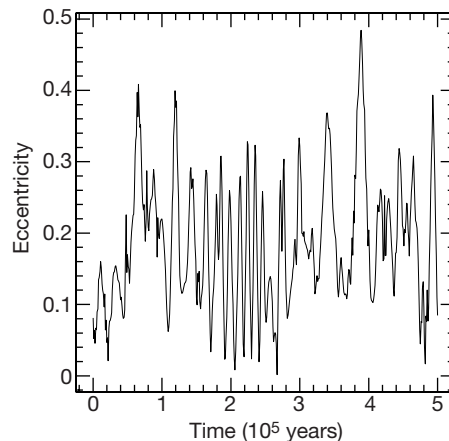


Figure 4 The orbital eccentricity of an object placed in the 3:1 resonance (at 2.5 AU) plotted as a function of time. The initial eccentricity is small, but chaotic perturbations from Jupiter force the eccentricity e of the asteroid to undergo a random walk, leading to a net increase in e and the eventual removal of the asteroid from the Solar System.

Box 1

Chaos and resonance overlap

Here we describe a simple model that captures many of the generic features of conservative or hamiltonian dynamical systems. As stated earlier, at least two oscillators must be interacting or coupled for chaos to occur. We can add a second oscillator to the rigid pendulum by vertically (or horizontally) moving the pivot of the pendulum periodically. Much insight can be gained by carefully examining this and similarly simple models^{4,6,27,28}.

To describe the motion we now need two angles, say the angle of the pendulum, θ , and the phase of the pivot, ψ . In addition we need two angle derivatives, $\dot{\theta}$ or $p = \dot{\theta}/\beta$, and $\dot{\psi}$. The hamiltonian is

$$H = \frac{1}{2}\beta p^2 + \alpha l + [\epsilon + 2K\cos\psi]\cos\theta$$

$$= \frac{1}{2}\beta p^2 + \alpha l + \epsilon\cos\theta + K\cos(\theta - \psi) + K\cos(\theta + \psi) \quad (1)$$

The first line presents an amplitude-modulated pendulum; the second presents three phase-modulated pendulums. See ref. 25 for a detailed development. Several constants depend on the physical characteristics of the driven pendulum: β is related to the moment of inertia of the pendulum; $2K$ is the amplitude of the driving; and ϵ is proportional to the gravitational acceleration. The constant α is the driving frequency or $\dot{\psi}$, and l is the negative of the total energy of the pendulum.

There are three resonances (cosine terms) which dominate the motion when their arguments are slowly varying: $\dot{\theta} \approx 0$, $\dot{\theta} \approx \dot{\psi}$, and $\dot{\theta} \approx -\dot{\psi}$. We can visualize the trajectories by plotting p versus θ whenever the driving reaches a specific phase, $\psi = 0$ for example. The resulting diagram is called a surface of section.

The figure shows surfaces of section for $\beta = 1$, $\epsilon = -1$, and $K = -0.3$. Panels **a**, **b** and **c** correspond to decreasing driving frequencies of $\alpha = 7$, 3 and 0.3, respectively. Here, the natural oscillation frequency of the pendulum is $\sqrt{|\epsilon|/\beta} = 1$. The resonant islands are located at $p = 0$ and $\beta p = \pm \dot{\psi} = \pm \alpha$. The centre resonant island has a half-width of $\Delta P = 2\sqrt{|\epsilon|/|\beta|} = 2$ and the other two have half-widths of $\Delta P = 2\sqrt{K/\beta} \approx 1.1$. So, for the highest driving frequency (**a**) the centres of the resonant 'islands' are separated by more than their widths. Aside from the additional resonant islands, two new features appear. First, two chains of secondary islands can be seen. Second, the separatrix of the main island has broadened into a fuzzy zone. This zone was traced out by a single trajectory. Each of the separatrices shows such a chaotic zone, although they are too small to discern. Nevertheless, what appear to be invariant curves corresponding to regular trajectories can still be seen. That both regular and chaotic trajectories should appear on the same surface of section is a generic feature of a divided phase space^{16,19,28}.

In **b** the driving frequency has been reduced, $\alpha = 3$, such that now the islands, ignoring their interaction, just touch each other. As the driving frequency approaches the natural oscillation frequency of the pendulum, the small chaotic zone seen in **a** envelopes the three resonant islands. This is the result of the beginning of resonance overlap²⁸.

In **c** the driving frequency has been further reduced, $\alpha = 0.3$, such

that now the resonances have widths much larger than their separations. The motion is chaotic near the separatrix, which appears as a single thick band surrounding a stable island, and is surrounded by invariant curves. As can be seen from the hamiltonian, the three resonant terms can also be viewed as a single resonance with an oscillating width, $\Delta P = 2\sqrt{|\epsilon| + 2K\cos\psi}/\beta$. The width ranges from $\Delta P \approx 1.3$ to $\Delta P \approx 2.5$, for the chosen values of ϵ and K . Here, the chaotic zone is roughly that region over which the separatrix sweeps. This is termed modulational chaos⁵¹ or adiabatic chaos⁴.

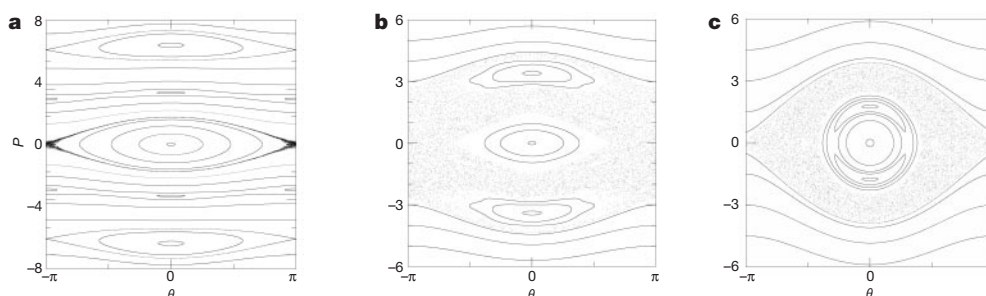
In discussing the origin of chaotic motion in the main text, we have treated the region of overlap as though it had no structure. This simplification is justified in some cases (for example in high-order resonances in the asteroid belt) and for some applications (for calculating Lyapunov times), but is not always adequate.

For example, the eccentricity of asteroids in the 3:1 resonance could librate around two different centres, one at the classical 'forced eccentricity' $e_f \approx 0.05$ and a second at $e \approx 0.15$ (ref. 3). The chaotic region surrounded the separatrix between these two points, allowing asteroids to pass from low (~ 0.05) to high (~ 0.3) eccentricities, where they could be removed by encounters with Mars. Later a third libration centre at even higher eccentricity was discovered, connected to the first two by chaotic orbits⁵². This allowed asteroids to reach $e > 0.06$; such objects are removed from the resonance, typically by plunging into the Sun. This motion is made possible by the resonance overlap, but describing it naturally requires knowledge of the underlying phase space structure; or we can perform numerical simulations.

Direct brute-force numerical integration indicates that plunging into the Sun is the most likely fate of bodies injected into the 3:1 resonance, with the remainder being pushed beyond Saturn⁵³. These integrations find a removal time of about one million years. There is also direct observational evidence that this interpretation is correct because the boundaries of the chaotic zone closely match the boundaries to observed distribution of asteroids³. Similar statements can be made about the other Kirkwood gaps and gaps in the outer asteroid belt⁶ (see Fig. 6).

A more realistic model for the 3:1 mean motion resonance, still for a single perturbing planet, would include terms of order I^2 , and allow for (different) I dependences in the factors K multiplying the cosine terms involving ψ . Such a model reproduces the two libration centres in the ψ , I motion found by Wisdom. Because the analytic model that we, and Wisdom, employ is of low order in e , it misses the third libration centre found in ref. 52.

More complete dynamical models, in particular those that include the effect of Saturn on Jupiter's orbit, show that secular resonances, that is, resonances between the precession periods of the asteroid and those of Jupiter and Saturn, overlap inside low-order mean motion resonances such as the 2:1^{54,55}. This overlap can enhance the rate of diffusion. Even this is not the entire story; Jupiter and Saturn are near (but not in) a 5:2 resonance. This 'great inequality' enhances the rate of diffusion of asteroids in the 2:1 mean motion resonance with Jupiter⁵⁶.



Box 1 Figure Surface of section for the driven pendulum. **a**, Driving frequency much faster than the natural pendulum oscillation frequency. **b**, Driving frequency approaching the natural pendulum

oscillation frequency. **c**, Driving frequency less than the natural pendulum oscillation frequency.

were destroyed in the vicinity of the 2:1 resonance¹. This removed the barrier of the KAM theorem. Unfortunately, these $\sim 100,000$ yr integrations¹, although they established the presence of chaotic orbits, did not give any hint as to how the irregular motion led to the removal of an asteroid from the resonance^{1,24}.

Then Wisdom showed by direct numerical integration that the eccentricity of small bodies placed in the 3:1 gap alternates chaotically between periods of low and moderate eccentricity as the result of perturbations from Jupiter² (see Fig. 4). This result relied upon the development of an algebraic mapping that could efficiently follow the long-term motion of asteroids in the 3:1 mean motion resonance with Jupiter for a million years or longer. This is a recurring theme in Solar System dynamics: more efficient algorithms and faster computers permit longer, more accurate integrations that often reveal unexpected dynamical results.

What is the source of the chaos in the 2:1 and 3:1 mean motion resonances? It has long been known that each mean motion resonance is composed of several individual resonant terms²⁵. For example in the 3:1 resonance there are three such terms in the planar case, as mentioned above. All of these resonances are weak, but they are also close together^{2-4,26}. It is the close proximity of multiple, albeit weak, resonances which invalidates the simple estimate for the integrated force given above. A single resonance produces a small force integrated over a libration period, and a simple 'cat's eye' in phase space with a smooth separatrix. Adding a second weak resonance breaks up this smooth separatrix in a chaotic region (see Box 1). The separatrix is broken because the second resonance 'nudges' the pendulum represented by the first resonance, altering the motion, particularly that of orbits that pass near the unstable fixed point²⁷. As a result the motion from one passage near the fixed point to the next is uncorrelated. The resonant force is still small compared to the force exerted by the Sun, and produces only small changes over a libration period. But over times longer than a libration period the asteroid experiences uncorrelated forces.

As a result of these uncorrelated forces the motion in the chaotic zone differs from that in a regular region; the chaotic motion is very sensitive to initial conditions. Suppose we take two copies of our interacting oscillators, released from slightly different positions. In one case we start near a stable fixed point, well away from the separatrix. Here the motion is regular and the two copies of the apparatus separate from each other linearly (or polynomially) in time (see Fig. 5). In the other case we start the two copies near an unstable fixed point. Now the separation increases exponentially in time. The timescale for this separation is called the Lyapunov time, T_L .

We can estimate the Lyapunov time of an asteroid placed artificially in the 3:1 mean motion resonance. There are only two relevant timescales in the problem: the libration period and the precession period. From our argument above, large-scale chaos occurs when the separation of the 'islands' (or 'cat's eyes') is comparable to their widths²⁸, that is, when the resonances overlap. Another way to say this is that large-scale chaos occurs when the libration and precession periods of a resonant asteroid are similar. Then the Lyapunov time must be equal to the libration period—the only timescale in the problem²⁹. In the case of the 3:1 resonance the precession period is 3×10^4 years, whereas the numerically determined Lyapunov time is 1.4×10^4 years.

The two types of trajectories differ in more than their rates of separation. The chaotic trajectory can explore a larger volume of phase space than is accessible by a regular trajectory. For asteroids this extra measure of freedom can be dangerous, opening routes by which they can be ejected or cannibalized by planets²⁻⁵ (it's world-eat-world out there).

In higher-order resonances, where the phase-space structure is relatively simple, the uncorrelated or chaotic forces lead to simple random walks involving only small exchanges of energy between the planet and the asteroid, because the semimajor axes of both bodies

are nearly fixed by the resonance condition. This is not so for the angular momentum; the torques associated with the uncorrelated forces produce a random walk in the angular momentum of both bodies. This leads to a classic gambler's ruin problem with the asteroid playing the part of the gambler and the planet taking the role of the house. If either body loses most of its angular momentum, that body's orbit becomes highly eccentric and subject to collisions with other objects in the Solar System or ejection—which is analogous to bankruptcy. The planet, with its much larger mass, has more capital (angular momentum) than the asteroid. Even worse, the asteroid is like a gambler forced to limit his winnings to what he can carry in his pockets; because of its small mass the asteroid cannot absorb enough angular momentum to produce a substantial change in the orbit of the planet. The two trade angular momentum back and forth, but when the inevitable happens and the asteroid loses its small stock of angular momentum, it is abruptly escorted from the cosmic casino.

We can estimate the time needed to remove the asteroid. Because the motion is chaotic we can treat the resonance angle as a random variable, which drives random changes in the eccentricity. In the case of the 3:1 resonance the changes in e^2 are of order $\Delta e^2 \approx (T_L/T_J)(M_J/M_\odot)e^2$, using the Lyapunov time, T_L , as the interval associated with the random changes (M_J , T_J , and $M_\odot$ are Jupiter's mass, Jupiter's orbital period, and the Sun's mass, respectively). The short-term average eccentricity diffuses to large values in a time of about one million years.

In some cases, such as the asteroid Helga in the 12:7 resonance, the motion is chaotic with a very short Lyapunov time, but the diffusion time is comparable to or larger than the age of the Solar System. When applied to Helga this theory predicts that the asteroid should survive for about 8 billion years. Direct numerical integrations agree with this prediction⁶.

We note that the largest gap, that between the 1:1 resonance harbouring the Trojan asteroids, and the Hildas in the 3:2 resonance, is the result of the overlap of distinct mean motion resonances³⁰; for example the 4:3 resonance overlaps with the 5:4, the 5:4 with the 6:5, and so forth.

It is believed that most meteorites come from one of two sources, the 3:1 mean motion resonance and the ν_6 secular resonance^{4,31,32}. The latter involves a resonance between the precession frequency of the apsidal line of the asteroid and the sixth fundamental secular frequency of the Solar System (which is very roughly Saturn's precession frequency). Roughly equal numbers of meteorites come from each type of resonance. Until recently there was one difficulty with this idea. The cosmic ray exposure ages (a measure of the delivery time) of the stony meteorites are typically 20 Myr (ref. 33), whereas the delivery time from the 3:1 resonance is much shorter, 1 Myr. However, recent simulations³⁴ suggest a way out of this dilemma: most meteorites, which are produced by collisions between larger bodies in the asteroid belt, are not injected directly into either the 3:1 or the ν_6 resonance. Rather, they are placed in the vicinity of the resonance, and then dragged into the unstable region by the Yarkovsky effect³⁴, which arises due to the anisotropic thermal radiation from those regions of the fragment which are exposed to the Sun. With the addition of this slow dynamical precursor, the theory of meteorite delivery appears to be complete.

Chaotic spin-orbit resonances

One of the most important examples of chaos is afforded by the evolution of planetary spins. This chaos is produced by resonances between spin and orbit precession periods. In such resonances, the asphericity of the planet couples to the non-axisymmetric perturbation produced by orbital eccentricity or inclination. The small deviations from perfect sphericity, typically a few parts in a thousand for larger bodies like planets, leads to significant exchanges of energy and angular momentum between orbital and

rotational (or spin) motions of satellites and even planets. Resonances between spin period and orbital period are common in the Solar System, the Moon being a prominent example. The phase space occupied by such resonances is small, making it unlikely that so many bodies formed in resonance. This paradox is explained by the fact that dissipative effects tend to drive bodies into these resonances, where the motion is stabilized^{35,36}. As a result, most resonant satellites are currently deep in their respective one-to-one spin-orbit resonances, where the motion is regular; however, in one case, Saturn's satellite Hyperion, the chaotic motion may have persisted until the present⁸.

Because solar tides are so weak, dissipative effects tend to be less important for planetary spins, allowing for richer dynamics. Ward³⁷ showed that the angle between the spin and orbital axes of Mars (the obliquity) varies by $\pm 13.6^\circ$ around its average of 24° over millions of years. These variations are a result of a resonance between the precession of the spin axis and a combination of orbital precession frequencies; improvements in orbital models resulted in an increased variation of $\pm 20^\circ$ (ref. 9). Numerical integrations then showed that the obliquity of Mars is evolving chaotically, and varies over an even larger range^{10,11}. Such a variation has profound, but as yet poorly understood, implications for climate variation. It is likely that the spin axes of Mercury and Venus underwent chaotic variations in the past¹¹.

The tilt of the Earth, currently 23° , will also increase in the future¹²; the Moon will evolve outward under the influence of the tides, resulting in a decrease in the precession rate of the Earth. Eventually the precession rate will become resonant with yet another combination of orbital precession rates. Once again, numerical integrations show that the Earth's obliquity will vary chaotically^{13,14}. The tilt of the Earth's axis may increase to 90° . The effect on our climate is hard to estimate, but the result is unlikely to be pleasant.

Three-body resonances

Numerical integrations of main-belt asteroids show that a substantial fraction of these bodies have chaotic orbits with rather short Lyapunov times ($\sim 10^5$ years)^{6,38,39}. Most of this chaotic motion is not associated with any two-body resonance. Instead it is the result of the interaction among three bodies: the asteroid, Jupiter, and either Mars in the inner asteroid belt or Saturn in the main or outer asteroid belt^{38,39}. As with Helga, the Lyapunov times can be quite short, but the diffusion times are comparable to or longer than the age of the Solar System.

Three-body resonances arise when one planet, such as Jupiter, is perturbed by a second, such as Saturn. The orbit of Jupiter is then no longer a simple keplerian ellipse. The potential experienced by the asteroid in Jupiter's gravitational field is given by an expression formally equivalent to the two-body case, but Jupiter's orbital elements now vary with time. This variation introduces a whole new suite of frequencies into the potential experienced by the asteroid; in addition to all the harmonics of Jupiter's period, all the harmonics of Saturn's orbital period appear as well. Three-body resonances have also been considered for the galilean and uranian satellites, as well as in ring systems^{40–42}.

These new terms in the potential have much smaller amplitudes than do two-body resonant terms that have the same number of powers of eccentricity. Three-body resonances are proportional to the product of the masses of the two perturbing bodies, M_J and M_S in our example, rather than just M_J as in a two-body resonance.

The structure of a three-body resonance is similar to that of a two-body resonance—multiple, very narrow components separated in semimajor axis by an amount that depends on the precession frequencies of the bodies involved. When the separation is comparable to or smaller than the width of the individual resonances, the motion in the immediate vicinity of the resonance will be chaotic. We can then estimate the Lyapunov time and the

diffusion time in the same manner as for two-body resonances.

We have searched the catalogue of asteroids for evidence of gaps at the location of a number of the stronger three-body resonances. We have not been able to identify a candidate gap.

We can check the theory in other ways. For example, some asteroids on the inner edge of the belt (near Mars) could have formed in three-body resonances involving the asteroid, Mars and Jupiter. If the diffusion time of some of these asteroids is comparable to the age of the Solar System, we should be able to find objects that are about to be removed from the belt owing to close encounters, or possibly collisions, with either Mars or Earth. Such bodies may have already been discovered: the so-called near-Earth asteroids⁴³. These bodies have been a puzzle for astronomers, as in their present orbits they have very short lifetimes, typically millions of years. Because the Solar System is billions of years old, objects with lifetimes of millions of years should have vanished long ago, unless there is some way to replenish the supply. The theory of three-body resonances offers a possible mechanism.

Chaos among the giant planets

Even trajectories outside but near a resonance can be affected by its presence. In the century following Newton's publication of his law of universal gravitation, astronomers noted that the positions of Jupiter and Saturn deviated from their predicted positions by some 30 minutes of arc. The difference became known as the great inequality. Laplace noted that the predictions did not take into account the influence of Saturn on Jupiter's orbit, and vice versa. On the face of it this seemed reasonable, because the mass of either planet was less than one one-thousandth that of the Sun. Early astronomers had made the naive estimate outlined above in our discussion of the Kirkwood gaps, which integrates this tiny force over the interval between successive conjunctions of the planets, a time somewhat longer than T_J . But Laplace realized that the orbital period of Saturn was almost exactly $5/2$ times that of Jupiter ($2/T_J - 5/T_S \approx 1/85T_J$). The perturbations accumulate over a much longer interval ($85T_J$), permitting a larger exchange of energy and angular momentum between the two planets. The change in the predicted position of the planet on the sky was roughly 85×85 times larger than the simple estimate would indicate. With this result, Laplace was able to reconcile the observations with the prediction of the law of gravity.

This discovery strongly affected Laplace's views regarding determinism, reflected in his well known statement⁴⁴:

The present state of the system of nature is evidently a consequence of what it was in the preceding moment, and if we conceive of an intelligence that at a given instant comprehends all the relations of the entities of this universe, it could state the respective position, motions, and general affects of all these entities at any time in the past or future.

Physical astronomy, the branch of knowledge that does the greatest honour to the human mind, gives us an idea, albeit imperfect, of what such an intelligence would be.

This view of the world passed into common currency: the clockwork motion of the planets became the epitome of regularity.

This view is wrong. It is wrong because it ignores the delicate nature of the separatrix, the cradle of chaos. The giant planets provide further evidence of the power of three-body resonances, the subtlety of natural phenomena, and the difficulty of interpreting the fruits of scientific enquiry.

The *tour de force* numerical integrations of the outer planets in 1988 (ref. 45) shattered the clockwork. Using the 'Digital Orrery', a parallel computer built specifically for the task and now part of the Smithsonian collection in Washington DC, they followed the motion of the four giant planets and Pluto (as a test particle) over 845 Myr. To the surprise of all at the time, Pluto's orbit was chaotic with $T_L \approx 10$ million years.

The breakthrough was followed by a re-examination of the orbital evolution of the full Solar System. Longer integrations^{46,47} of an approximate model of the Solar System (the orbits were averaged, and Pluto was ignored) showed that the Solar System itself was chaotic. Laskar suggested that secular resonances involving the terrestrial planets were responsible for the chaotic motion^{46,47}. Subsequent integrations of a complete model¹⁵ confirmed that the full Solar System was chaotic with $T_L \approx 5$ Myr. Sussman and Wisdom also found that the four giant planets by themselves appeared to form a chaotic system¹⁵.

The chaos amongst the four giant planets arises from a three-body resonance involving Jupiter, Saturn and Uranus¹⁶. The orbital period of Uranus is nearly seven times that of Jupiter. In terms of the orbital frequency $n = 2\pi/T$, the difference $n_J - 7n_U$ is equal to $5n_S - 2n_J$. The relevant resonant terms in the potential experienced by Uranus are proportional to $M_J M_S / (2 - 5n_S/n_J)$. This ‘small denominator’, $(2 - 5n_S/n_J) \approx 1/85$, enhances the potential experienced by Uranus by a factor of 85. It was of course this small denominator that Laplace used to explain the origin of the great inequality, and which is implicated in the removal of objects from the 2:1 resonance in the asteroid belt.

The width of a single component of this three-body resonance is tiny, comparable to the radius of Uranus, but so is the separation between components; the resonances just overlap. The predicted Lyapunov time is $T_L \approx 10$ million years. Because Uranus is so much less massive than either Jupiter or Saturn, its orbital angular momentum is substantially less than that of the two gas giants. By the analogy above, it is hence subject to the gambler’s ruin. The diffusion time (roughly the time before Uranus is ejected) is 10^{18} years, much longer than the current age of the Universe. The theory also predicts the location and Lyapunov times of other chaotic zones near the present orbit of Uranus. Detailed numerical integrations verify the presence and Lyapunov times of all these zones, as well as illustrating the transitions between libration and rotation in the relevant resonances. Current computational resources are inadequate to test the predicted diffusion time.

Chaos among the terrestrial planets

As noted above, numerical integrations of the full Solar System, inner planets included, show evidence of chaos with a Lyapunov time of $T_L \approx 5$ million years^{15,17}. Unlike the outer planets, the source of this chaos has not been convincingly established. Laskar pointed to what are called “secular resonances” between the terrestrial planets as a candidate source of the chaos⁴⁶. He

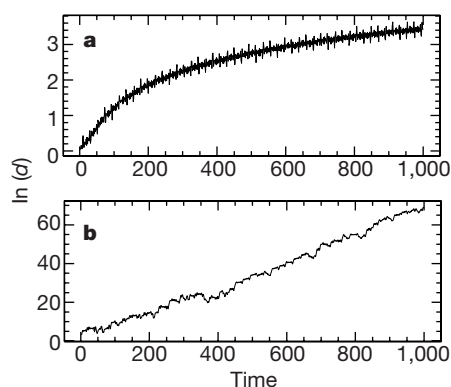


Figure 5 The distance between two initially nearly identical initial conditions for two interacting nonlinear oscillators. **a**, The motion is near a stable fixed point; **b**, the motion starts near an unstable fixed point. In **b** (but not in **a**) the two initial conditions separate exponentially with time, and are said to be chaotic.

found what appears to be an alternation between circulation and libration in the angles $\sigma_1 \equiv (\bar{\omega}_1 - \bar{\omega}_5) - (\Omega_1 - \Omega_2)$ and $\sigma_2 \equiv 2(\bar{\omega}_4 - \bar{\omega}_3) - (\Omega_4 - \Omega_3)$. In these relations $\bar{\omega}_1$ refers to the orientation of Mercury’s apsidal line, which is the line from the Sun to the point of Mercury’s orbit closest to the Sun, whereas Ω_1 refers to the orientation of Mercury’s nodal line; the latter is defined by the intersection of the orbital plane of Mercury with the orbital plane of the Earth. Similar definitions apply to the elements for Venus (subscript 2), Earth (3), Mars (4), and Jupiter (5). However, these two resonances do not interact directly with each other, so by themselves they are unlikely to produce large-scale chaos. Later integrations⁴⁷ identified a third resonance, $\sigma_3 \equiv (\bar{\omega}_4 - \bar{\omega}_3) - (\Omega_4 - \Omega_3)$. Laskar found that it librated while σ_2 rotated, and vice versa. As σ_2 and σ_3 involve the same degrees of freedom, they are a more promising candidate for overlapping resonances. Sussman and Wisdom¹⁵, employing a more realistic (unaveraged) model, confirmed the alternate libration and rotation of σ_2 , but never saw σ_3 librate; this does not rule out the possibility that the overlap of the associated resonances causes the chaos. However, their assessment that “no dynamical mechanism for the observed chaotic behaviour of the Solar System has been clearly demonstrated” seems warranted, at least for the terrestrial planets. Without a clear identification of the source of the chaos it is not possible to use an analytic development, such as was used for the outer planets, to confirm the Lyapunov time and then estimate the timescale for diffusion of the system.

Chaos in other planetary systems

By analogy with our Solar System, multiplanet systems seem likely. In fact three Jupiter-mass objects orbit Upsilon Andromeda⁴⁸. Any multiple-planet system is subject to the instabilities described in this Review. All of the known planetary systems orbiting solar-type stars have ages of 10^9 years or greater. This, together with the copernican assumption that we are not observing the system at a privileged time, such as immediately after a recent planetary ejection, can be used to put limits on the masses and orbital elements of the planets, quantities which cannot be tightly constrained or even directly observed using current techniques^{49,50}. Dynamical constraints can

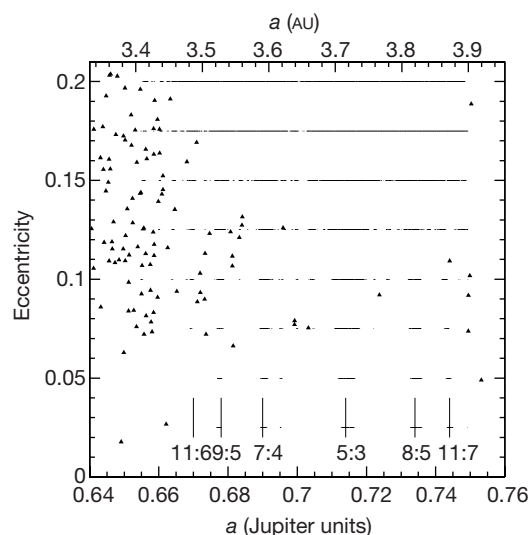


Figure 6 The location of asteroids in the outer asteroid belt in the (eccentricity, semimajor axis) plane. Lyapunov times have been computed as a function of the semimajor axis for several values of eccentricity. Points are plotted when the Lyapunov time is below a threshold value. These indicate the chaotic zones. It is clear that the known asteroids avoid these regions.

also be placed on the existence of smaller (Earth-mass) bodies in orbits near those of the Jupiter-mass objects. Observational searches for such low-mass companions will have to await improved techniques. □

1. Giffen, R. A study of commensurable motion in the asteroid belt. *Astron. Astrophys* **23**, 387–403 (1973).
2. Wisdom, J. The origin of the Kirkwood gaps: a mapping for asteroidal motion near the 3/1 commensurability. *Astron. J.* **87**, 577–593 (1982).
3. Wisdom, J. Chaotic behavior and the origin of the 3/1 Kirkwood gap. *Icarus* **56**, 51–74 (1983).
4. Wisdom, J. A perturbative treatment of motion near the 3/1 commensurability. *Icarus* **63**, 272–289 (1985).
5. Wisdom, J. Meteorites may follow a chaotic route to earth. *Nature* **315**, 731–733 (1985).
6. Holman, M. & Murray, N. Chaos in high-order mean motion resonances in the outer asteroid belt. *Astron. J.* **112**, 1278–1293 (1996).
7. Malhotra, R., Duncan, M. J. & Levison, H. F. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 1231 (Univ. Arizona Press, Tucson, 2000).
8. Wisdom, J., Peale, S. J. & Mignard, F. The chaotic rotation of Hyperion. *Icarus* **58**, 137–152 (1984).
9. Ward, W. R. & Rudy, D. J. Resonant obliquity of Mars? *Icarus* **94**, 160–164 (1991).
10. Touma, J. & Wisdom, J. The chaotic obliquity of Mars. *Science* **259**, 1294–1297 (1993).
11. Laskar, J. & Robutel, P. The chaotic obliquity of the planets. *Nature* **361**, 608–612 (1993).
12. Ward, W. R. Comments on the long-term stability of the earth's obliquity. *Icarus* **50**, 444–448 (1982).
13. Laskar, J., Joutel, F. & Robutel, P. Stabilization of the earth's obliquity by the moon. *Nature* **361**, 615–617 (1993).
14. Neron de Surgy, O. & Laskar, J. On the long term evolution of the spin of the Earth. *Astron. Astrophys.* **318**, 975–989 (1997).
15. Sussman, G. J. & Wisdom, J. Chaotic evolution of the solar system. *Science* **257**, 56–62 (1992).
16. Murray, N. & Holman, M. The origin of chaos in the outer solar system. *Science* **283**, 1877–1881 (1999).
17. Laskar, J. A numerical experiment on the chaotic behaviour of the solar system. *Nature* **338**, 237–238 (1989).
18. Chirikov, B. V. Resonance processes in magnetic traps. *Plasma Phys.* **1**, 253–260 (1960).
19. Hénon, M. & Heiles, C. The applicability of the third integral of motion: Some numerical experiments. *Astron. J.* **69**, 73–79 (1964).
20. Kirkwood, D. *Meteoritic Astronomy: A Treatise on Shooting-stars, Fireballs, and Aerolites* 105–111 (Lippincott, Philadelphia, 1867).
21. Kolmogorov, A. N. On conservation of conditionally periodic motions under small perturbations of the Hamiltonian. *Dokl. Akad. Nauk* **98**, 527–530 (1953).
22. Arnold, V. I. Proof of A. N. Kolmogorov's theorem on the preservation of quasi-periodic motions under small perturbations of the Hamiltonian. *Uspehi Mat. Nauk.* **18**, 13–40 (1963); English translation *Russian Math. Surv.* **18**, 9–36 (1963).
23. Moser, J. On invariant curves of area-preserving mappings of an annulus. *Nachr. Akad. Wiss. Göttingen Math. Phys. K* **1**, 87–120 (1962).
24. Froeschle, C. & Scholl, H. The stochasticity of peculiar orbits in the 2/1 Kirkwood gap. *Astron. Astrophys.* **93**, 62–66 (1981).
25. Peirce, B. Development of the perturbative function of planetary motion. *Astron. J.* **1**, 1–8 (1849).
26. Goldreich, P. in *Cosmology and Astrophysics* (eds Terzian, Y. & Bilson, E. M.) 121–129 (Cornell Univ. Press, Ithaca, 1982).
27. Lichtenberg, A. J. & Lieberman, M. A. *Regular and Chaotic Dynamics* 245–264 (Springer, New York, 1992).
28. Chirikov, B. V. A universal instability of many-dimensional oscillator systems. *Phys. Rep.* **52**, 263–379 (1979).
29. Murray, N. & Holman, M. Diffusive chaos in the outer asteroid belt. *Astron. J.* **114**, 1246–1259 (1997).
30. Wisdom, J. The resonance overlap criterion and the onset of stochastic behavior in the restricted three-body. *Astron. J.* **85**, 1122–1133 (1980).
31. Wetherill, G. W. Where do the Apollo objects come from? *Icarus* **76**, 1–18 (1988).
32. Morbidelli, A. & Gladman, B. Orbital and temporal distributions of meteorites originating in the asteroid belt. *Meteor. Planet. Sci.* **33**, 999–1016 (1998).
33. Marti, K. & Graf, Cosmic-ray exposure history of ordinary chondrites. *Annu. Rev. Earth Planet. Sci.* **20**, 221–243 (1992).
34. Vokrouhlický, D. & Farinella, P. Efficient delivery of meteorites to the Earth from a wide range of asteroid parent bodies. *Nature* **407**, 606–608 (2000).
35. Darwin, G. H. On the precession of a viscous spheroid and on the remote history of the Earth. *Phil. Trans. R. Soc. Lond.* **170**, 447–538 (1879).
36. Goldreich, P. Final spin states of planets and satellites. *Astron. J.* **71**, 1–7 (1966).
37. Ward, W. R. Large-scale variations in the obliquity of Mars. *Science* **181**, 260–262 (1973).
38. Murray, N., Holman, M. & Potter, M. On the origin of chaos in the asteroid belt. *Astron. J.* **116**, 2583–2589 (1998).
39. Nesvorný, D. & Morbidelli, A. Three-body mean motion resonances and the chaotic structure of the asteroid belt. *Astron. J.* **116**, 3029–3037 (1998).
40. Tisserand, F. *Traité de Mécanique Céleste* Tome IV (Gauthier-Villars, Paris, 1896).
41. Goldreich, P. & Nicholson, P. Revenge of tiny Miranda. *Nature* **269**, 783–785 (1977).
42. Aksnes, K. in *Planetary Rings* (ed. Brahic, A.) 479–487 (Proceedings of IAU Colloq. 75, Cepadues, Toulouse, 1984).
43. Morbidelli, A. & Nesvorný, D. Numerous weak resonances drive asteroids toward terrestrial planet's orbits. *Icarus* **139**, 295–308 (1999).
44. Laplace, P. S. *Oeuvres Complètes de Laplace* Vol. 8, 144–145; (Imprimerie Royale, Paris, 1843–47); translation Gillispie, C. C., Fox, R. & Grattan-Guinness, I. *Pierre-Simon Laplace* (Princeton Univ. Press, 1997).
45. Sussman, G. J. & Wisdom, J. Numerical evidence that the motion of Pluto is chaotic. *Science* **241**, 433–437 (1988).
46. Laskar, J. The chaotic motion of the solar system—A numerical estimate of the size of the chaotic zones. *Icarus* **88**, 266–291 (1990).
47. Laskar, J. in *Proc. IAU Symposium 152* (ed. Ferraz-Mello, S.) 1–16 (1990).
48. Butler, R. P. *et al.* Evidence for multiple companions to ν Andromedae. *Astrophys. J.* **526**, 916–927 (1999).
49. Laughlin, G. & Adams, F. C. Stability and chaos in the ν Andromedae planetary system. *Astrophys. J.* **526**, 881–889 (1999).
50. Rivera, E. J. & Lissauer, J. J. Stability analysis of the planetary system orbiting ν Andromedae. *Astrophys. J.* **530**, 454–463 (2000).
51. Tennyson, J. in *Nonlinear Dynamics in Beam-Beam Interactions* (eds Month, M. & Herrera, J. C.) 158–193 (AIP Conf. Proc. 57, AIP, New York, 1979).
52. Ferraz-Mello, S. & Klafke, J. C. in *Predictability, Stability, and Chaos in N-body Dynamical Systems* (ed. Roy, A. E.) 177–184 (Plenum, New York, 1991).
53. Gladman, B. J. *et al.* Dynamical lifetimes of objects injected into asteroid belt resonances. *Science* **277**, 197–201 (1997).
54. Henrard, J. & Lemaitre, A. A perturbation method for problems with two critical arguments. *Celest. Mech.* **39**, 213–238 (1986).
55. Morbidelli, A. & Moons, M. Secular resonances in mean motion commensurabilities: the 2:1 and 3:2 cases. *Icarus* **102**, 316–332 (1993).
56. Ferraz-Mello, S., Michtchenko, T. A. & Roig, F. The determinant role of Jupiter's great inequality in the depletion of the Hecuba gap. *Astron. J.* **116**, 1491–1500 (1998).

Acknowledgements

We thank P. Goldreich for helpful conversations. This research was supported by NSERC of Canada and NASA.

Correspondence and requests for materials should be addressed to M.H. (e-mail: mholman@cfa.harvard.edu).

Crystal structure of a hairpin ribozyme–inhibitor complex with implications for catalysis

Peter B. Rupert & Adrian R. Ferré-D'Amaré

Division of Basic Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109-1024, USA

The hairpin ribozyme catalyses sequence-specific cleavage of RNA. The active site of this natural RNA results from the docking of two irregular helices: stems A and B. One strand of stem A harbours the scissile bond. The 2.4 Å resolution structure of a hairpin ribozyme–inhibitor complex reveals that the ribozyme aligns the 2'-OH nucleophile and the 5'-oxo leaving group by twisting apart the nucleotides that flank the scissile phosphate. The base of the nucleotide preceding the cleavage site is stacked within stem A; the next nucleotide, a conserved guanine, is extruded from stem A and accommodated by a highly complementary pocket in the minor groove of stem B. Metal ions are absent from the active site. The bases of four conserved purines are positioned potentially to serve as acid-base catalysts. This is the first structure determination of a fully assembled ribozyme active site that catalyses a phosphodiester cleavage without recourse to metal ions.

The hairpin ribozyme is a catalytic RNA derived from the self-cleaving and ligating domain of the negative polarity strand of the satellite RNA of tobacco ringspot virus. *In vivo*, this domain is responsible for generating unit-length circular satellite RNA during the course of its rolling-circle replication (reviewed in ref. 1). The cleavage reaction generates products with 5'-hydroxyl and 2',3'-cyclic phosphate termini (Fig. 1a), which are analogous to those produced by three other natural ribozymes that are part of circular satellite RNAs: the hammerhead, the hepatitis delta virus (HDV) and the Varkud satellite (VS) ribozymes. These four ribozymes, however, are structurally unrelated, and therefore represent independent evolutionary solutions to the same biochemical problem (reviewed in ref. 2). None of these ribozymes require proteins for activity. They are also quite small, for example the hairpin ribozyme requires about 50 nucleotides for activity. The hairpin ribozyme is an ideal experimental system to help understand how a small catalytic RNA self-assembles and forms an active site.

Biochemical experiments demonstrated that the active site of the hairpin ribozyme results from the association of two largely helical segments of the satellite RNA, stems A (one strand of which contains the scissile phosphate) and B (Fig. 1b). Both contain nucleotides necessary for catalysis¹. The stems can be synthesized as two separate RNAs and mixed *in vitro* to reconstitute an active ribozyme^{3,4}. Most biochemical experiments have been carried out using constructs where the two stems are connected by a single-stranded linker. In the satellite RNA, stems A and B are part of a four-helix junction⁵. Fluorescence resonance energy transfer (FRET) measurements demonstrated that the docking of stems A and B is greatly favoured in constructs that contain the four-helix junction, compared with systems where the two stems are connected by an extended linker^{5,6}. Docking is facilitated by the cooperative binding of two molar equivalents of alkaline earth metal ions, which have dissociation constants of 10^{-10} to 10^{-9} M (refs 5, 7); however, the hairpin ribozyme does not use metal ions to activate its 2'-OH nucleophile. Moreover, it has been shown to be fully active with cobalt (III) hexammine^{8–10}, which mimics hydrated Mg^{2+} , but cannot make direct coordination with water or RNA ligands¹¹. Thus, the RNA itself provides functional groups that accomplish catalysis, unlike 'classical' ribozymes such as the hammerhead or the group I intron that use metal ion cofactors¹. Evidence suggests that nucleotide bases with perturbed pK_a values

in the active sites of the HDV ribozyme^{12–14} and the peptidyl-transferase of the 50S ribosomal subunit¹⁵ enable those RNAs to carry out general acid-base catalysis. Such a mechanism is potentially used by the hairpin ribozyme.

As a step towards elucidating its catalytic mechanism, and to provide a structural framework for rationalizing the large body of biochemical information gathered on this catalytic RNA, we have now determined the structure of a fully assembled hairpin ribozyme (Table 1 and Methods). We crystallized a four-helix junction form of the RNA in complex with a substrate strand in which the nucleophilic 2'-OH group (Fig. 1a) was replaced with a methoxy function (Fig. 1b). Such modified substrates are competitive inhibitors of the ribozyme^{16,17}. Crystals were grown in the presence of 20 mM Ca^{2+} , enough to saturate the structural metal-ion-binding sites. The cognate site for the RNA-binding protein UIA was grafted onto the structurally and functionally dispensable¹ distal end of stem B, and the engineered RNA was crystallized together with the basic protein (see Methods and Supplementary Information). The structure consists of two independent hairpin ribozyme–inhibitor complexes in different crystalline environments. Although we did not

Table 1 Crystallographic statistics

Data set	Crystal I			Crystal II
	$\lambda 1$	$\lambda 2$	$\lambda 3$	
Diffraction data				
Resolution (Å)	30.0–2.9	30.0–2.9	30.0–2.9	100.0–2.4
Wavelength (Å)	0.9796	0.9793	0.9566	1.1000
Reflections				
Total/unique*	83,395/44,497	168,452/45,575	168,844/45,520	148,993/41,290
Completeness (%)	98.2 (89.6)	99.1 (95.0)	99.1 (94.9)	95.6 (73.9)
$\langle I \rangle / \langle \sigma(I) \rangle$	19.7 (1.7)	26.5 (2.3)	24.6 (2.2)	22.8 (2.1)
$R_{\text{sym}} (\%)^\dagger$	6.4 (41.0)	8.0 (42.3)	7.7 (43.7)	6.7 (26.1)
Phasing statistics				
Phasing power‡	1.26 (0.24)	1.34 (0.36)	1.10 (0.27)	
$R_{\text{Krat}}§$	0.05 (0.20)	0.04 (0.15)	0.04 (0.15)	

Numbers in parentheses refer to the high-resolution shell (3.0–2.9 Å for crystal I; 2.5–2.4 Å for crystal II).

* For crystal I data, the Bijvoet pairs were not merged.

† $R_{\text{sym}} = \sum |I - \langle I \rangle| / \sum I$ where I is the observed intensity and $\langle I \rangle$ is the statistically weighted absolute intensity of multiple measurements of symmetry related reflections.

‡ Phasing power = $\langle |F_o| \rangle / \langle |F_c| \rangle$, reported for all acentric reflections.

§ $R_{\text{Krat}} = \sum (|F_o| - |F_c|) / \sum |F_o|$, reported for all reflections.

use non-crystallographic symmetry (NCS) restraints, the structures of the two catalytic cores have a root-mean-square (r.m.s.) difference of $< 0.5 \text{ \AA}$ for all atoms, which is comparable to the precision of the atomic coordinates (see Methods). The significant agreement

of the refined crystallographical model with the results of biochemical experiments, and the in-line-attack conformation present at the catalytic site, suggest that this $2.4\text{-\AA}$ resolution structure accurately depicts the hairpin ribozyme in an active conformation.

Overview of the structure

The four helical stems that comprise the hairpin ribozyme radiate from a perfectly base-paired four-way junction. In the crystal structure, the helices form two coaxial stacks: stem D stacks on A, and stem C stacks on B. The two stacks cross at around a 60° angle, and are connected by a pair of anti-parallel crossovers (Figs 1b and 2). The junction allows the central portions of stems A and B to dock through their minor grooves and form the active site of the ribozyme. No bound metal ions are observed in the junction. In the crystal structure of an unrelated four-helix junction formed by a DNA–RNA hybrid¹⁸, a bound metal ion was found in the major groove of one of the B-form helices. Because the stems of the hairpin ribozyme are in an A-form conformation near the junction, there is no equivalent binding site in its structure. Well-ordered metal ions are observed in the non-canonical moiety of stem B (see below). These could correspond to the cooperatively bound metal ions detected by FRET^{5,7}, as tight binding of divalent ions was detected only in experiments with the intact ribozyme, but not the isolated fully-base-paired junction¹⁹.

The stacking order was correctly predicted on the basis of gel-mobility shift and FRET experiments⁵. It was proposed that the end-to-end distances of the stems would be in the order $AB < CD < BD$, as is observed in the crystal structure (Fig. 2). The structure is consistent with the results of hydroxyl radical probing experiments, which found that a solvent-inaccessible core is formed when the RNA adopts its active structure in the presence of polyvalent cations²⁰. The ten C4' atoms found to be solvent-inaccessible in that study (G11–A14, C25–C27, A38 and U42–43) are fully buried in the structure. Although a model of the hairpin ribozyme generated on the basis of biochemical data and long-distance crosslinking experiments²¹ is substantially different from the crystal structure (the r.m.s. difference for 66 C1' atoms is $4.3 \text{ \AA}$), the orientation of stems A and B is the same.

Stems A and B of the hairpin ribozyme each consist of two A-form, Watson–Crick-paired segments flanking a stretch of nucleotides that adopt non-canonical conformations (Fig. 1b). The solution three-dimensional structures of isolated stems A and B have been determined^{22,23}. As these separate RNAs can be mixed *in vitro* to reconstitute a functional ribozyme^{3,4}, the nuclear magnetic resonance (NMR) structures probably represent the conformation of the two stems before docking. Comparison with the crystal structure shows that large conformational rearrangements take place in both RNA stems on assembly of the active site.

Stem A conformation

In the docked state, the non-canonical moiety of stem A consists of three single-hydrogen-bond base pairs and two unpaired nucleotides (Fig. 3a). Of the latter, the conserved, essential nucleotide following the scissile phosphate, G+1 (see Fig. 1 legend for numbering scheme), is extruded from the helical stack, and adopts a syn conformation (Fig. 3b) as proposed previously²⁴. The hairpin ribozyme tolerates any base at position –1 (ref. 24). In the crystal structure, the N3 atom of A–1 accepts a single hydrogen bond from A9. This base-pair geometry is compatible with all four bases, as both purines have an N3, and both pyrimidines have an O2 atom in the equivalent position. The geometry of the A–1·A9 pair causes A–1 to make a cross-strand stack onto G8. This and the conformation of G+1 result in a pronounced broadening of the minor groove (Fig. 3a).

The three non-canonical base pairs observed in the crystal structure are absent in the isolated stem A. In the NMR structure, an anti conformation G+1 is accommodated in the helical stack (Fig. 3c), and forms a sheared pair with A9 (ref. 22). The NMR

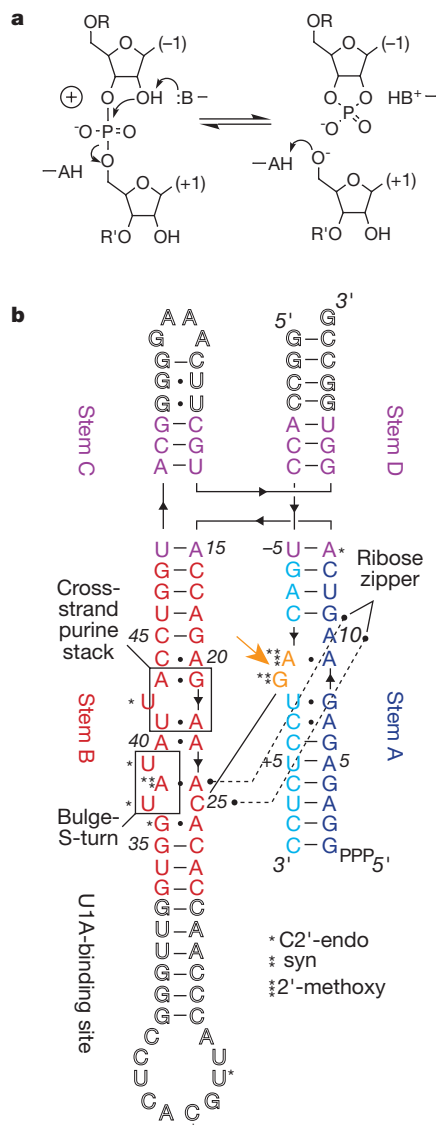


Figure 1 Sequence and reactions catalysed by the hairpin ribozyme. **a**, Reversible transesterification reaction catalysed by the ribozyme. Hypothetical acid, base and positive charge catalytic groups are indicated. **b**, Sequence and secondary structure of the crystallization construct. The RNA consists of a 92-nucleotide ribozyme and a 21-nucleotide substrate strand. In the latter, the 2'-OH adjacent to the scissile bond was substituted with a methoxy to inhibit cleavage. To facilitate crystallization²⁷, a cognate site for the RNA-binding protein U1A was grafted at the distal end of stem B. This construct was decided on after screening 21 different RNA–U1A complexes. The RNA is numbered according to convention¹. In the substrate strand, numbers of nucleotides in the 5' and 3' direction of the cleavage site (yellow arrow) are preceded by minus and plus signs, respectively. Ribozyme-strand nucleotides are numbered starting at the 5' terminus, skipping the grafted U1A-binding site. Lines with embedded arrows indicate the direction of the chains at the crossovers. Solid letters, wild-type sequence; outline letters, nucleotides added for crystallization. Thin lines indicate Watson–Crick pairs; black circles indicate non-canonical pairs involving at least one hydrogen bond between co-planar bases. Nucleotides that participate in a ribose zipper²⁵ are connected by dotted lines. Nucleotides with C2'-endo ribose pucker (asterisk), those in the syn conformation (double asterisk), and the A–1 residue that carries the 2'-methoxy modification (triple asterisk) are indicated. The same nucleotide conformations are present in both molecules in the crystallographic asymmetric unit.

construct has a C base at position -1, and this pairs with a protonated A10. The latter pair is inconsistent with the lack of a requirement for a particular base at -1 (ref. 22). The minor groove width of the stem A conformation observed by NMR is considerably narrower than that of the docked state (Fig. 3c).

Stem B conformation

In the crystal structure, stem B contains six non-canonical base pairs that stack in an irregular underwound helix (Fig. 4a). A single-hydrogen-bond A20·C44 pair is followed by the sheared G21·A43 and reverse-Hoogsteen U41·A22 pairs. The latter two pairs conform closely to the cross-strand purine stack motif²⁵. In the crystal structure, U42 is extruded from the helical stack, and its base lies perpendicular to the plane of the helical base pairs. This arrangement is consistent with the observation that a photo-crosslink can be formed at high efficiency between G21 and/or A22 and U41. A crosslink to U42 was observed only at high concentrations where the ribozyme RNA formed a non-functional dimer²⁶.

The bottom of the non-canonical moiety of stem B is stretched apart by an N1–N1 pair between G36 and A26 (Fig. 4a). This is followed by an S-turn²⁷, where the direction of the ribose sugars in the backbone is inverted twice, resulting in extrusion of the U37 base into solvent. A38 lies within the stack in a syn conformation, and makes a single hydrogen bond with A24. The S-turn is followed by a second bulged pyrimidine, U39. The bulging of U39 is consistent with the observation that it can be replaced with a propyl linker without affecting ribozyme activity²⁸. The combination of an S-turn with a bulged U and a cross-strand A-stack results in a broad minor groove into which the bases of A38 and U42 protrude, to participate in the interhelical interface.

Concurrent with the wide minor groove of stem B is a narrow major groove that seems to be stabilized by the coordination of several divalent metal ions. Two of these are particularly well ordered, and engage in both inner-sphere and outer-sphere coordi-

nation (Figs 2a and 4a). In our crystals, their anomalous scattering identifies them unequivocally as Ca^{2+} (see Methods). Either calcium ion is compatible with a non-selective metal-ion-binding site identified by backbone hydrolysis with terbium (III) (ref. 29), as well as a cobalt-hexammine-binding site identified by NMR studies of the free stem B (ref. 30). Cobalt hexammine can only make outer-sphere contacts¹¹. As the calcium ions in the crystal structure make several inner-sphere contacts, this implies that the conformation of the RNA backbone in the neighbourhood of the bulged S-turn is somewhat flexible.

The conformation of the isolated stem B (ref. 23) differs greatly from that seen in the crystal structure. Only two of the seven non-canonical base pairs observed by NMR (A20·C44 and G21·A43) are retained in the docked conformation. Whereas U41 is extruded into an S-turn in the undocked stem B, it forms a base pair with A22 in the crystal structure. Superposition of the two structures (Fig. 4b, c) demonstrates the pronounced widening of the minor groove of stem B that takes place on docking.

The rearrangements of stems A and B, involving dramatic changes in base-pairing schemes and significant alteration of backbone geometry concomitant with formation of tertiary structure, are not unprecedented. For instance, NMR studies of the P5abc domain of the *Tetrahymena* group I intron have demonstrated that local changes in secondary structure take place in that RNA as it undergoes Mg^{2+} -dependent folding³¹. In contrast, when Mg^{2+} is added to the isolated stems A and B of the hairpin ribozyme in solution, neither undergoes a detectable change in conformation^{22,30}. Thus, binding energy resulting from assembly of the active site, not metal ion coordination, initiates the structural rearrangements described above.

Interhelical interface

The assembly of the hairpin ribozyme buries 1,570 Å² of solvent-accessible surface area of the minor-groove faces of the cores of

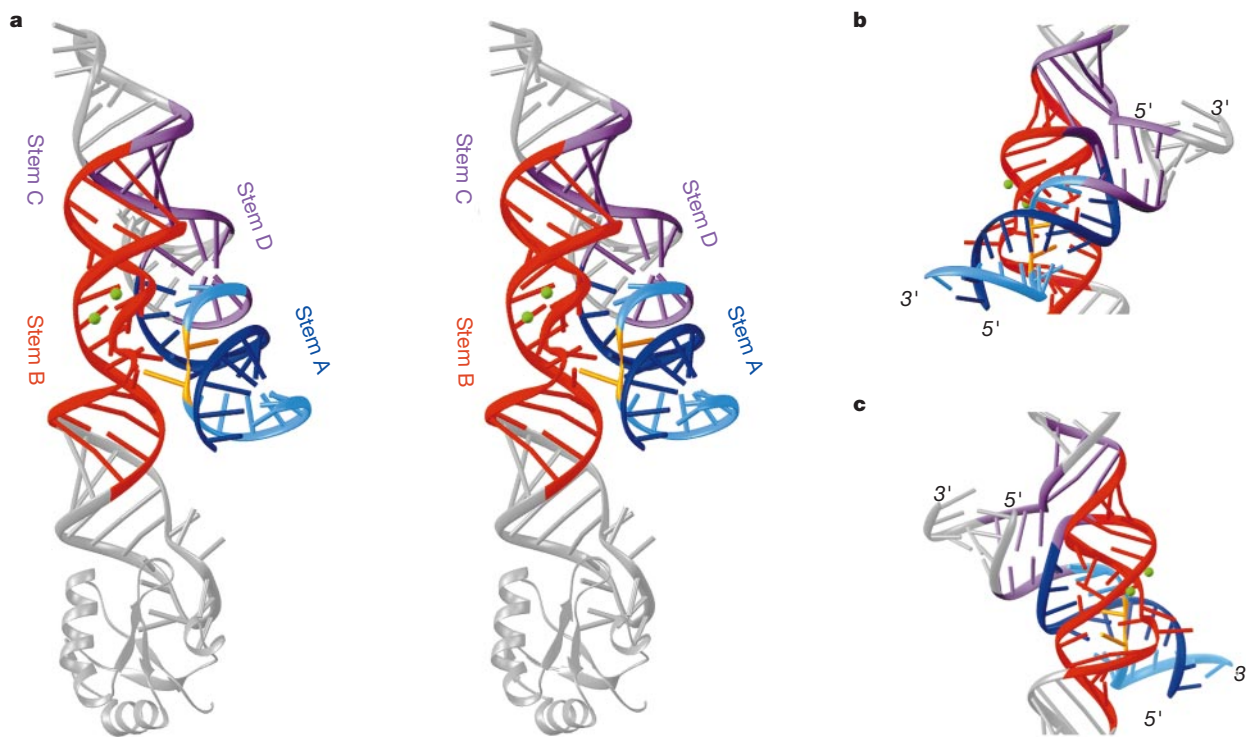


Figure 2 Architecture of the hairpin ribozyme. The backbone of the two RNA strands is depicted as ribbons; the nucleotide bases as sticks (colour code as in Fig. 1). Two tightly bound calcium ions are represented as green spheres. **a**, Stereoview of the docked conformation of the hairpin ribozyme. The scissile bond lies between the two yellow

nucleotides, which are splayed apart. The U1A protein is shown as a grey ribbon.

b, c, Views orthogonal to that in **a**, showing the path of the RNA chains in the four-helix junction, the crossing angle between stems A and B, and the approximately 30° bend at the site of the bulged nucleotides in stem B.

stems A and B (see Methods). A conventional ribose zipper^{12,25} stitches together the N3 atoms of A10 and A24, and the 2'-OHs of nucleotides 10, 11, 24 and 25 (Fig. 5a). Because these four 2'-OH groups were found to be important for ribozyme activity^{17,32,33}, two different proposals for ribose zippers have been made^{21,33}, neither of which predicted the correct pattern of hydrogen bonds.

U42, which is extruded from the stem B stack (Fig. 4a), packs in the interface (Fig. 5a). Its base makes five hydrogen bonds with functional groups of nucleotides in both stem A and B. Previous studies have shown that abolishing these hydrogen bonds is deleterious to ribozyme activity. For example, methylation of the N6 atom of A22 and A23, methylation of the N2 of G11 and deletion of the 2'-OH of U12 all result in reduced activity³³. Methylation of O2 of U41 reduces activity to 0.06% of the wild type³⁴. These and other modification experiments were carried out using minimal hairpin ribozymes; the effects on activity may be less pronounced for the more stable construct of the four-helix junction.

The docked conformation of stem B (Fig. 4a) creates a pocket that binds the extruded syn G+1 nucleotide of stem A (Fig. 3a). Docking of the two stems through their broad minor grooves enables G+1 to make a Watson–Crick pair with C25 (Fig. 5a, b). This was correctly

predicted on the basis of compensatory mutations³⁵. A proposed³⁵ base triple between G+1, C25 and A9 is not observed. A38, which is also in the syn conformation, and A26 form the roof and floor of the G-binding pocket. In addition to these stacking interactions, G+1 donates a hydrogen bond from its 2'-OH to G36, and receives a hydrogen bond to its N7 from the 2'-OH of A38 (Fig. 5b). Thus, the pocket is highly complementary to the chemical identity of G. Individual mutations of G+1, C25, or A38 to any other nucleotide abolish ribozyme activity^{36,37}. Substitutions that would destroy some of the hydrogen bonds in which G+1 participates, such as its replacement for inosine or methylation of O6, result in a complete loss of activity³⁸.

Active site and catalysis

Consistent with biochemical observations that the hairpin ribozyme does not require direct coordination to metal ions for catalysis^{8–10}, the crystal structure shows that the active site consists entirely of RNA. Neither MAD nor model-phased electron density maps show features that could correspond to tightly-bound metal ions in the vicinity of the scissile bond (see Supplementary Information Figs 2 and 3). The extrusion of G+1 from the helical stack of stem A, presumably stabilized by its burial into its cognate pocket in stem B, constrains the riboses of both A–1 and G+1 into C2'-endo puckers (B-form-like twists of the furanose rings). The conformation of these two nucleotides results in an in-line arrangement of the nucleophile (2'-OH of A–1) and leaving group (5'-oxo of G+1) relative to the phosphorus atom at the reaction site (Fig. 5d), as is expected for this ribozyme as it approaches the transition state of its

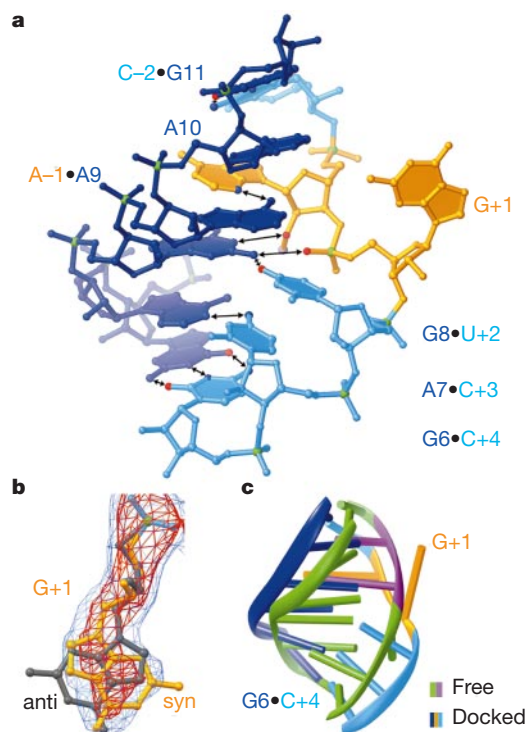


Figure 3 Structure of stem A. **a**, Conformation of nucleotides in the core of stem A in the docked ribozyme, seen from the minor groove. In this and subsequent figures phosphorus atoms (green), and oxygen and nitrogen atoms participating in noteworthy hydrogen bonding (two-headed arrows) are highlighted (red and blue, respectively). The extrusion of G+1 from the helical stack results in the A–1 torsion angles ϵ and ζ being +anticlinal (*ac*) and –antiperiplanar (*ap*), and the G+1 torsion angles α , β and γ being –synperiplanar (*sp*), –*ac* and –synclinal (*sc*). For comparison, the five angles are –*ac*, –*sc*, –*sc*, –*ap* and +*sc* in A-form RNA⁴⁵. **b**, The 2.4-Å resolution ‘solvent-flattened’ experimental electron density map corresponding to nucleotide G+1 is represented as blue and red mesh, for contour levels at one and two standard deviations above mean peak height. G+1 adopts an unusual syn conformation (yellow) in both molecules in the asymmetric unit. For comparison, the incorrect anti conformation is shown (grey). **c**, Superposition of the docked and undocked²² structures of stem A. The free state is shown in green, with nucleotides flanking the scissile bond in magenta. Note the difference in G+1 conformation. The minor groove is considerably narrower in the undocked RNA. The r.m.s. difference between the two structures (for twelve C1' atoms) is 3.6 Å.

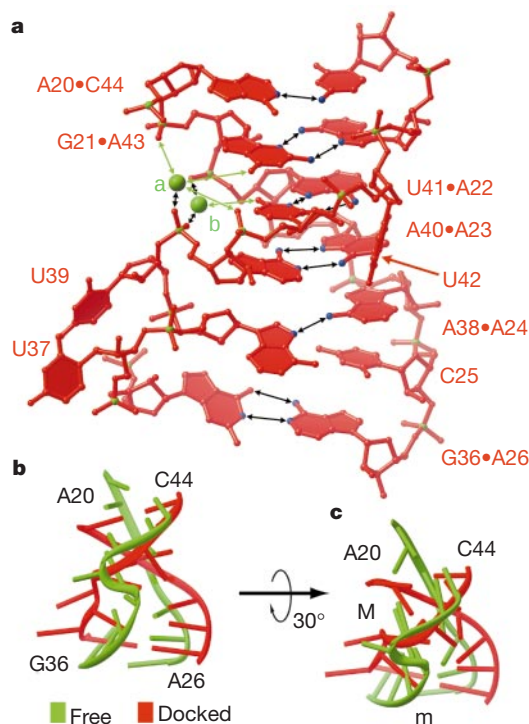


Figure 4 Structure of Stem B. **a**, Docked conformation of stem B, seen from its minor groove. For the major-groove calcium ions, inner-sphere coordination (~ 2.5 Å) is indicated by black arrows; outer-sphere coordination (~ 4.5 Å) by green arrows. The Ca^{2+} ion labelled *b* is very prominent in MAD and anomalous difference Fourier maps (4.3 and 3.8 s.d. above mean peak height). **b**, Comparison of the docked and undocked²³ conformations. The least-squares superposition of the two structures (red and green ribbons, respectively) yields a r.m.s. difference of 6.0 Å (for the 16 C1' atoms shown in **a**). **c**, View resulting from tilting the top of **b** toward the viewer. Upper and lower case ‘m’ denote the major and minor grooves, respectively. Note how the minor groove broadens and the major groove narrows on docking.

S_N2 (substitution nucleophilic, bimolecular) transesterification reaction³⁹. The in-line arrangement is observed in both molecules in the asymmetric unit. The nucleophilic 2' oxygen atom is blocked by a methyl group in our crystallization construct. Even at 2.4 Å resolution, the position of this methyl is defined in both MAD and omit maps (Fig. 5c), establishing unambiguously the RNA backbone conformation near the scissile bond.

The splayed substrate strand conformation that is observed in the active site of the hairpin ribozyme is markedly similar to that of the dinucleotide mimic UpcA, when bound to ribonuclease-A (ref. 40). In this structure, His 12 lies close to the nucleophile (2'-OH of U) and serves as a general base. The imidazole ring of His 119 stacks on the adenine base, and lies next to the leaving group (5'-oxo of A), serving as a general acid^{41,42}. Lysine 41 counteracts the negative charge that develops in the transition state⁴⁰. Which RNA functional groups could carry out catalysis in the hairpin ribozyme? In the crystal structure, G8 hydrogen bonds to both the pro-Rp oxygen of the scissile phosphate (through N2: 3.0 and 3.1 Å in the two independent complexes in the asymmetric unit) and the nucleophilic 2'-OH (through N1: 2.9 and 2.7 Å, Figs 3a and 5b). The purine base of A38 stacks on G+1. It does not participate in hydrogen bonding in the active site of this ribozyme-inhibitor complex, but if the 5'-oxo leaving group were to swing out, it would be within hydrogen-bonding distance of N1 of A38 (the C5' of G+1 is 3.4 Å from N1 of A38 in both molecules). Neither G8 nor A38 can be mutated to any other nucleotide without abolishing hairpin ribozyme activity^{36,37}. Replacement of G8 by inosine or 6-methyl-G impairs ribozyme activity^{33,38}. A38 is held in place by hydrogen bonding to A24 (Fig. 5b). Deletion of the N6 and N7 involved in this interaction is deleterious^{33,38}. Two additional purines present in the vicinity of the scissile bond (Fig. 5d), A9 and A10, have been implicated in ribozyme activity by mutation and functional-group

substitution^{33,37,38}. In the active conformation, neither residue uses its N1 to make hydrogen bonds. Instead, the Watson–Crick faces of these adenines are in van der Waals contacts with surrounding residues. A cavity⁴³ large enough to hold a single water molecule exists here, but electron density maps show no evidence that it is occupied (not shown).

We suggest several catalytic functions for active-site residues that are consistent with the crystal structure and available biochemical data. In the ligation reaction, which the hairpin ribozyme favours over cleavage by a factor of about 30 (ref. 44), G8 and A38 could be functioning as a general acid and a general base. The positions of G8 and A38 relative to the scissile bond are very similar to those of His 12 and His 119, respectively, in ribonuclease-A. The A38–G+1 stacking interaction is markedly similar to the stacking of His 119 onto the A base of UpcA. The pK_a values of the N1 of guanosine and adenosine 5'-monophosphates are 10 and 3.9, respectively⁴⁵. Thus, a considerable change in acidity of these purines would be required for them to catalyse the reverse cleavage reaction in this manner. Alternatively, the N1 atoms of G8 and A38 could serve only to orient the nucleophilic 2'-OH and 5'-OH in the cleavage and ligation reactions, respectively. The hydrogen bond between N2 of G8 and the pro-Rp oxygen would help to relieve the developing negative charge on the scissile phosphate. If either A9 or A10 were to be protonated (as a result of a pK_a shift) they could have a function that is analogous to Lys 41 in ribonuclease-A. Such an electrostatic function would be compatible with their distance from the scissile phosphorus (the N1–P distances are 6.7 Å for A9 in both molecules in the asymmetric unit, and 8.2 and 8.4 Å for A10). Finally, A9 or A10 (with a shifted pK_a value) could activate a water molecule that would then deprotonate the nucleophilic 2'-OH in the cleavage reaction (or the 5'-OH in the ligation reaction). For this to happen, the ribozyme would have to 'breathe' to allow access to solvent.

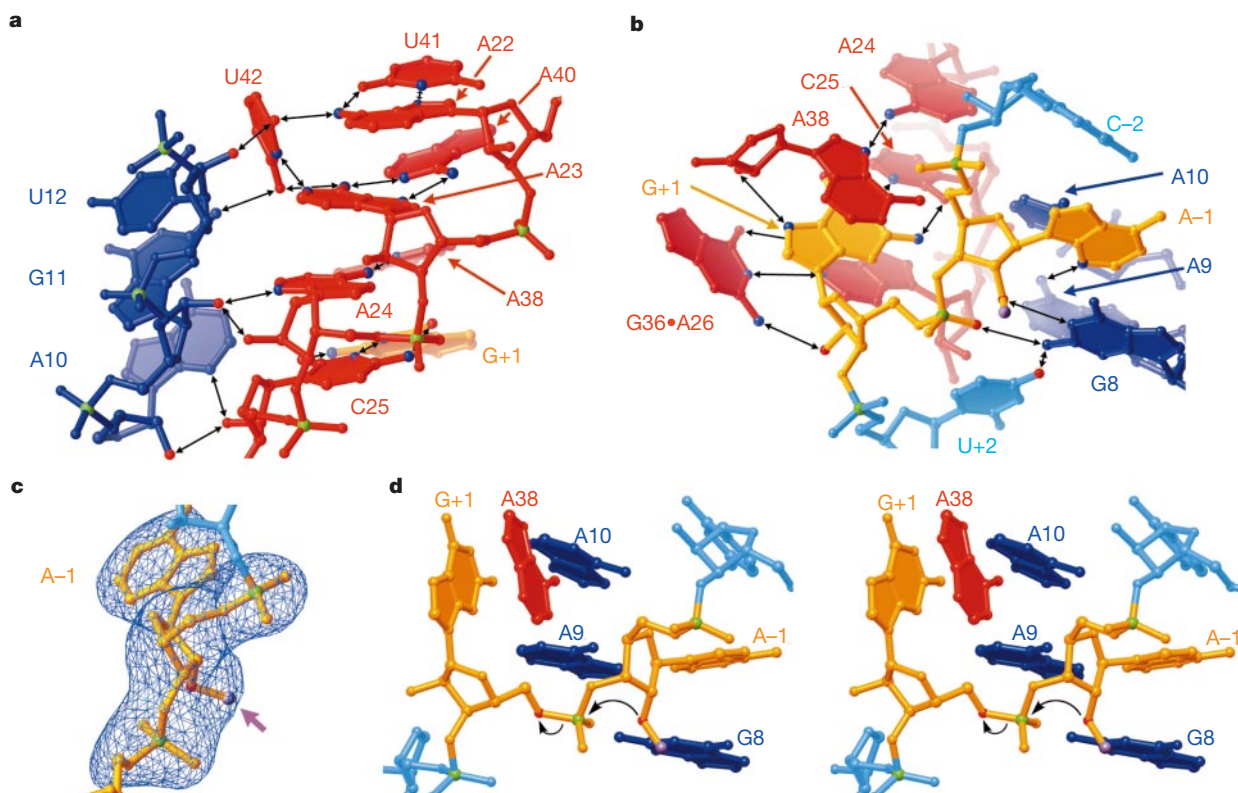


Figure 5 Active site architecture. **a**, Interface of stems A and B, seen from the side opposite to that of Fig. 2a. **b**, Detail of the G+1-binding pocket. **c**, Simulated-annealing $2F_o - 2F_c$ omit density map contoured around A-1 at one s.d. The arrow points to the bulge

in the density that corresponds to the methyl group blocking the 2' oxygen. **d**, Stereoview of the active site, showing the in-line-attack conformation and several purine bases that may participate in catalysis.

The crystal structure of a hairpin ribozyme–inhibitor complex demonstrates how an RNA that catalyses a transesterification can align the 2'-OH nucleophile and the 5'-oxo leaving group of the reaction, and suggests how this nucleic acid could be carrying out acid-base catalysis. This structure represents the starting point for further crystallographic and biophysical investigation of the mechanism of action of this natural catalytic RNA. Of particular interest will be the structural determination of a ribozyme–product complex, to understand how the ligation reaction is catalysed by the same active site, the elucidation of the roles of G8, A9, A10 and A38 in catalysis, and the measurement of their pK_a values. If, indeed, the acidity of these purines is shifted, it will raise an important question: how does the organization of the active site accomplish this perturbation? □

Methods

Crystallization and data collection

The insert of plasmid pHF16, which encodes the ribozyme studied followed by a VS cleavage site⁴⁶, was prepared by PCR from overlapping synthetic DNA oligonucleotides. RNA and U1A-RNA-binding domain (RBD) preparation have been described²⁷. The substrate strand RNA was from Dharmacon Research. Grafting the U1A-RBD-binding site onto the end of stem B, and binding of U1A-RBD, did not affect catalytic activity, as judged by comparison with wild-type constructs (Supplementary Information, Fig. 1). Ribozyme and substrate strands (0.5 mM each) were mixed in 2.5 mM MgCl₂, and annealed (slow cooling from 65–22 °C). RNA was mixed 1:1 with U1A-RBD to a final complex concentration of 125 μM in 625 μM MgCl₂. Drops consisting of 5 μl each of complex and reservoir solution (20 mM CaCl₂, 220–260 mM NH₄Cl and 21% MPD) were pre-equilibrated by vapour diffusion at 22 °C. We obtained single crystals by seeding these drops with washed fragments of polycrystalline clusters grown at higher MPD concentrations. Crystals grew to 600 × 300 × 300 μm³ in several weeks. Crystals were stabilized in 5 mM spermine, 20 mM CaCl₂, 200 mM NH₄Cl and 30% MPD, and flash cooled by plunging into liquid nitrogen. The crystals (space group C2) have unit cell parameters $a = 256.3$ Å, $b = 44.1$ Å, $c = 102.4$ Å and $\beta = 109^\circ$, contain two RNA–protein complexes per asymmetric unit, have a solvent content of 65%, and are stabilized by a combination of protein–protein (5% of solvent-accessible surface area buried in crystal contacts), protein–RNA (30%) and RNA–RNA (65%) interactions. All crystals contained a substrate strand with a 2'-methoxy substitution at residue A–1. Crystal I contained selenomethionine protein. Crystal II contained U1A-RBD with normal methionine, and a substrate strand with an additional 5-iodo-uracil substitution at position 5. Diffraction data (Table 1) were measured (MAD data with the inverse-beam method) at 100 K at beamline 5.0.2 of the Advanced Light Source (ALS, Lawrence Berkeley National Laboratory) and reduced with the HKL package⁴⁷. Data-sets with stronger iodine and calcium anomalous signals were collected with Cu K α radiation for crystal II, and a third crystal that contained 5-iodo-uracil at position 7 (not shown).

Phase determination and structural refinement

Six selenium sites were located in crystal I data with CNS⁴⁸, and confirmed by difference Patterson maps. Two additional sites were located in difference Fourier syntheses. Heavy atom parameters were refined and phases calculated at 2.9 Å resolution using MAD data from crystal I (Table 1). The mean overall figure of merit (FOM) was 0.42 (0.08 between 3.0 and 2.9 Å). Density modification and phase extension⁴⁸ to 2.4 Å resolution against amplitudes from crystal II produced an electron density map (Fig. 3b; see also Supplementary Information, Fig. 2), into which most of the macromolecule residues could be built unambiguously (overall FOM = 0.9; 0.99 between 2.5 and 2.4 Å) using program O (ref. 49). Correct registration of both of the protein chains was confirmed by the selenium positions. As the protein–RNA interface structure is known²⁷, this also confirmed the register of RNA residues in stem B. Anomalous difference Fourier syntheses with amplitudes from the crystals with 5-iodo-uracil at positions 5 and 7 in stem A confirmed the sequence register of this RNA moiety.

Rounds of manual rebuilding, interspersed with positional, torsion-angle-simulated annealing, and restrained individual B-factor refinement⁴⁸ (restraints adjusted to minimize R_{free}), produced the current model ($R_{\text{free}} = 28.9\%$, $R_{\text{work}} = 24.7\%$; $R_{\text{free}} = 41.6\%$ and $R_{\text{work}} = 40.6\%$ between 2.5 and 2.4 Å). Refinement was against all observed crystal II amplitudes (37,148 and 4,142 in the work and test sets) and crystal I MAD phase-probability distributions, using a maximum-likelihood target⁴⁸. A solvent mask and an overall B-factor correction were used throughout; NCS restraints were not. The model contains all RNA atoms for both copies of the ribozyme–inhibitor complex (except for both 5'-γ-phosphates, which are disordered), U1A-RBD residues 6–97 for both chains, 48 water molecules, and 17 calcium and 2 chloride ions (6,405 non-hydrogen atoms). The average B-factor for all RNA atoms in the catalytic core (defined as residues –5 to +3, 7–14, 21–27 and 35–43) is 80.1 Å². Despite this, electron density is well defined for the ribozyme core (Figs 3b and 5c; see also Supplementary Information, Figs 2 and 3). The average real-space R-factor (in a composite simulated-annealing omit map) for all RNA residues in the catalytic core is 6.7%. For comparison, the protein atoms have an average B-factor of 49.9 Å², and have an average real-space R-factor of 4.4%. Ribose puckers were restrained to either C2'-endo or C3'-endo to account for residual features in $|F_o| - |F_c|$ syntheses. There are no breaks in the backbone electron density of the ribozyme core at 0.8 s.d. in a

composite σ_A -weighted simulated-annealing omit $2|F_o| - |F_c|$ map. The cross-validated σ_A mean coordinate error is 0.51 Å. The current model has r.m.s deviations of 0.0114 Å and 1.41° from ideal bond lengths and angles. Of the protein residues, 89.9% lie in the most favoured regions of the Ramachandran plot, with the remainder in additional allowed regions. Figures 2–5 were prepared (using RIBBONS⁵⁰) with molecule B, which is the best-ordered in the ribozyme core. Stems C and D are better ordered in molecule A. The r.m.s. difference between the two complexes in the asymmetric unit is 1.62 Å for all macro-molecule atoms, and 0.47 Å for all RNA atoms in the core. A probe radius of 1.4 Å was used to measure solvent-accessible surfaces.

Received 21 December 2000; accepted 12 February 2001.

- Fedor, M. J. Structure and function of the hairpin ribozyme. *J. Mol. Biol.* **297**, 269–291 (2000).
- McKay, D. B. & Wedekind, J. E. in *The RNA World* (eds Gesteland, R. F., Cech, T. R. & Atkins, J. F.) 265–286 (Cold Spring Harbor Press, Cold Spring Harbor, 1999).
- Butcher, S. E., Heckman, J. E. & Burke, J. M. Reconstitution of hairpin ribozyme activity following separation of functional domains. *J. Biol. Chem.* **270**, 29648–29651 (1995).
- Shin, C. *et al.* The loop B domain is physically separable from the loop A domain in the hairpin ribozyme. *Nucleic Acids Res.* **24**, 2685–2689 (1996).
- Murchie, A. I. H., Thomson, J. B., Walter, F. & Lilley, D. M. J. Folding of the hairpin ribozyme in its natural conformation achieves close physical proximity of the loops. *Mol. Cell* **1**, 873–881 (1998).
- Walter, N. G., Burke, J. M. & Millar, D. P. Stability of hairpin ribozyme tertiary structure is governed by the interdomain junction. *Nature Struct. Biol.* **6**, 544–549 (1999).
- Walter, F., Murchie, A. I. H., Thomson, J. B. & Lilley, D. M. J. Structure and activity of the hairpin ribozyme in its natural junction conformation: effect of metal ions. *Biochemistry* **37**, 14195–14203 (1998).
- Hampel, A. & Cowan, J. A. A unique mechanism for RNA catalysis: the role of metal cofactors in hairpin ribozyme cleavage. *Chem. Biol.* **4**, 513–517 (1997).
- Nesbitt, S., Hegg, L. A. & Fedor, M. J. An unusual pH-independent and metal-ion independent mechanism for hairpin ribozyme catalysis. *Chem. Biol.* **4**, 619–630 (1997).
- Young, K. J., Gill, F. & Grasby, J. A. Metal ions play a passive role in the hairpin ribozyme catalyzed reaction. *Nucleic Acids Res.* **25**, 3760–3766 (1997).
- Cowan, J. A. Metallochemistry of RNA. $\text{Co}(\text{NH}_3)_6^{3+}$ as a probe for Mg^{2+} (aq) binding sites. *J. Inorg. Biochem.* **49**, 171–175 (1993).
- Ferré-D'Amaré, A. R., Zhou, K. & Doudna, J. A. Crystal structure of a hepatitis delta virus ribozyme. *Nature* **395**, 567–574 (1998).
- Perrotta, A. T., Shih, I. & Been, M. D. Imidazole rescue of a cytosine mutation in a self-cleaving ribozyme. *Science* **286**, 123–126 (1999).
- Nakano, S. -I., Chadalavada, D. M. & Bevilacqua, P. C. General acid-base catalysis in the mechanism of a hepatitis delta virus ribozyme. *Science* **287**, 1493–1497 (2000).
- Muth, G. W., Ortoleva-Donnelly, L. & Strobel, S. A. A single adenosine with a neutral pK_a in the ribosomal peptidyl transferase center. *Science* **289**, 947–950 (2000).
- Chowrira, B. M. & Burke, J. M. Binding and cleavage of nucleic acids by the “hairpin” ribozyme. *Biochemistry* **30**, 8515–8522 (1991).
- Walter, N. G., Hampel, K. J., Brown, K. M. & Burke, J. M. Tertiary structure formation in the hairpin ribozyme monitored by fluorescence resonance energy transfer. *EMBO J.* **17**, 2378–2391 (1998).
- Nowakowski, J., Shim, P. J., Prasad, G. S., Stout, C. D. & Joyce, G. F. Crystal structure of an 82-nucleotide RNA–DNA complex formed by the 10–23 DNA enzyme. *Nature Struct. Biol.* **6**, 151–156 (1999).
- Walter, F., Murchie, A. I. H. & Lilley, D. M. J. Folding of the four-way RNA junction of the hairpin ribozyme. *Biochemistry* **37**, 17629–17636 (1998).
- Hampel, K. J., Walter, N. G. & Burke, J. M. The solvent-protected core of the hairpin ribozyme–substrate complex. *Biochemistry* **37**, 14672–14682 (1998).
- Earnshaw, D. J. *et al.* Inter-domain cross-linking and molecular modeling of the hairpin ribozyme. *J. Mol. Biol.* **274**, 197–212 (1997).
- Cai, Z. & Tinoco, I. J. Solution structure of loop A from the hairpin ribozyme from tobacco ringspot virus satellite. *Biochemistry* **35**, 6026–6036 (1996).
- Butcher, S. E., Allain, F. H. -T. & Feigon, J. Solution structure of the loop B domain from the hairpin ribozyme. *Nature Struct. Biol.* **6**, 212–216 (1999).
- Chowrira, B. M., Berzal-Herranz, A. & Burke, J. M. Novel guanosine requirement for catalysis by the hairpin ribozyme. *Nature* **354**, 320–322 (1991).
- Moore, P. B. Structural motifs in RNA. *Annu. Rev. Biochem.* **68**, 287–300 (1999).
- Butcher, S. E. & Burke, J. M. A photo-cross-linkable tertiary structure motif found in functionally distinct RNA molecules is essential for catalytic function of the hairpin ribozyme. *Biochemistry* **33**, 992–999 (1994).
- Ferré-D'Amaré, A. R. & Doudna, J. A. Crystallization and structure determination of a hepatitis delta virus ribozyme: use of the RNA-binding protein U1A as a crystallization module. *J. Mol. Biol.* **295**, 541–556 (2000).
- Schmidt, S. *et al.* Base and sugar requirements for RNA cleavage of essential nucleoside residues in internal loop B of the hairpin ribozyme: implications for secondary structure. *Nucleic Acids Res.* **24**, 573–581 (1996).
- Walter, N. G., Yang, N. & Burke, J. M. Probing non-selective cation binding in the hairpin ribozyme with Tb(III). *J. Mol. Biol.* **298**, 539–555 (2000).
- Butcher, S. E., Allain, F. H. -T. & Feigon, J. Determination of metal ion binding sites within the hairpin ribozyme domains by NMR. *Biochemistry* **39**, 2174–2184 (2000).
- Wu, M. & Tinoco, I. RNA folding causes secondary structure rearrangement. *Proc. Natl Acad. Sci. USA* **95**, 11555–11560 (1998).
- Chowrira, B., Berzal-Herranz, A., Keller, C. F. & Burke, J. M. Four ribose 2'-hydroxyl groups essential for catalytic function of the hairpin ribozyme. *J. Biol. Chem.* **268**, 19458–19462 (1993).
- Ryder, S. P. & Strobel, S. A. Nucleotide analog interference mapping of the hairpin ribozyme: implications for secondary and tertiary structure formation. *J. Mol. Biol.* **291**, 295–311 (1999).
- Young, K. J. *et al.* The role of essential pyrimidines in the hairpin ribozyme-catalysed reaction. *J. Mol. Biol.* **288**, 853–866 (1999).
- Pinard, R. *et al.* Structural basis for the guanosine requirement of the hairpin ribozyme. *Biochemistry* **38**, 16035–16039 (1999).

36. Siwkowski, A., Shippy, R. & Hampel, A. Analysis of hairpin ribozyme base mutations in loops 2 and 4 and their effects on cis-cleavage *in vitro*. *Biochemistry* **36**, 3930–3940 (1997).
37. Shippy, R., Siwkowski, A. & Hampel, A. Mutational analysis of loops 1 and 5 of the hairpin ribozyme. *Biochemistry* **37**, 564–570 (1998).
38. Grasy, J. A., Mersmann, K., Singh, M. & Gait, M. J. Purine functional groups in essential residues of the hairpin ribozyme required for catalytic cleavage of RNA. *Biochemistry* **34**, 4068–4076 (1995).
39. van Tol, H., Buzayan, J. M., Feldstein, P. A., Eckstein, F. & Bruening, G. Two autolytic processing reactions of a satellite RNA proceed with inversion of configuration. *Nucleic Acids Res.* **18**, 1971–1975 (1990).
40. Richards, F. M. *et al.* Protein structure, ribonuclease-S, and nucleotide interactions. *Cold Spring Harbor Symp. Quant. Biol.* **36**, 35–43 (1971).
41. Findlay, D., Herries, D. G., Mathias, A. P., Rabin, B. R. & Ross, C. A. The active site and mechanism of action of bovine pancreatic ribonuclease 7. The catalytic mechanism. *Biochem. J.* **85**, 152–153 (1962).
42. Roberts, G. C. K., Dennis, E. A., Meadows, D. H., Cohen, J. S. & Jardetzky, O. The mechanism of action of ribonuclease. *Proc. Natl Acad. Sci. USA* **62**, 1151 (1969).
43. Kleywegt, G. J. & Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D* **50**, 175–177 (1994).
44. Fedor, M. J. Tertiary structure stabilization promotes hairpin ribozyme ligation. *Biochemistry* **38**, 11040–11050 (1999).
45. Saenger, W. *Principles of Nucleic Acid Structure* (Springer, New York, 1984).
46. Ferré-D'Amaré, A. R. & Doudna, J. A. Use of *cis*- and *trans*-ribozymes to remove 5' and 3' heterogeneities from milligrams of *in vitro* transcribed RNA. *Nucleic Acids Res.* **24**, 977–978 (1996).
47. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
48. Brünger, A. T. *et al.* Crystallography and NMR system: a new software system for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
49. Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
50. Carson, M. Ribbons. *Methods Enzymol.* **277**, 493–505 (1997).

Supplementary information is available on *Nature's* World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of *Nature*.

Acknowledgements

We thank T. Earnest, L. Hung and G. McDermott for help at ALS beamline 5.0.2; C. Hoang for biochemical support; J. Bolduc, P. Heath and B. Shen for computational and crystallographic support; K. Nagai for U1A plasmids; and S. Biggins, M. Holmes, P. Li, M. Rosenberg, S. Ryder, S. Sigurdsson, B. Stoddard, S. Strobel, R. Strong, G. Varani, D. Wilson and K. Zhang for discussions. This work was supported by institutional funds from the Fred Hutchinson Cancer Research Center (FHCRC). Access to ALS beamline 5.0.2 as part of the principal research consortium was made possible by general support from the FHCRC. P.B.R. is a post-doctoral trainee of the Chromosome Metabolism and Cancer training grant from the National Cancer Institute to the FHCRC.

Correspondence and requests for materials should be addressed to A.R.F. (e-mail: aferr@fhcrc.org). Atomic coordinates have been deposited in the Protein DataBank under accession number 1HP6.

An auroral flare at Jupiter

J. H. Waite Jr*, G. R. Gladstone†, W. S. Lewis‡, R. Goldstein†, D. J. McComas†, P. Riley‡, R. J. Walker§, P. Robertson†, S. Desai||, J. T. Clarke* & D. T. Young*

* Department of Atmospheric, Oceanic, and Space Sciences, University of Michigan, Ann Arbor, Michigan 48109, USA

† Southwest Research Institute, PO Drawer 28510, San Antonio, Texas 78228, USA

‡ SAIC, 10260 Campus Point Drive, San Diego, California 92121, USA

§ Institute of Geophysics and Planetary Physics, UCLA, Los Angeles, California 90095, USA

|| 1305 Addiewell Place, San Jose, California 95120, USA

Jupiter's aurora is the most powerful in the Solar System¹. It is powered largely by energy extracted from planetary rotation², although there seems also to be a contribution from the solar wind^{3,4}. This contrasts with Earth's aurora, which is generated through the interaction of the solar wind with the magnetosphere. The major features of Jupiter's aurora (based on far-ultraviolet^{5–7}, near-infrared^{8,9} and visible-wavelength¹⁰ observations) include a main oval that generally corotates with the planet and a region of patchy, diffuse emission inside the oval on Jupiter's dusk side. Here we report the discovery of a rapidly evolving, very bright and localized emission poleward of the northern main oval, in a region connected magnetically to Jupiter's outer magnetosphere. The intensity of the emission increased by a factor of 30 within 70 s, and then decreased on a similar timescale, all captured during a

single four-minute exposure. This type of flaring emission has not previously been reported for Jupiter (similar, but smaller, transient events have been observed at Earth), and it may be related directly to changes in the solar wind.

Four imaging observations of Jupiter's northern far-ultraviolet aurora were performed with the Hubble Space Telescope Imaging Spectrograph (STIS) in time-tagged mode during a roughly two-hour period on 21 September 1999 (Fig. 1). Time-tagged images obtained during the second observation, between 21:04:50 and 21:08:50 UT, reveal a dramatic, rapidly intensifying, flare-like auroral emission at jovicentric latitudes between 60° and 70° (Fig. 2). These latitudes lie poleward of the main oval, which generally coincides with the ionospheric footprint of magnetic field lines that map to corotating regions in the magnetic equatorial plane; the outer boundary of these regions is thought to lie near 30 R_J where R_J is the radius of Jupiter^{5–7}. The event begins as a small, 'pinpoint' emission near 167° System III longitude and 63°N (frames 1–3), which rises rapidly in intensity and becomes a structure several thousand kilometres across (frames 5–12). The event reaches its maximum intensity at ~70 s (frame 8), after which it starts to decrease in both intensity and size (frames 9–14 and 19–24). (STIS stopped counting temporarily during frames 15–18 owing to a full buffer. The light curve of the flare is well represented by an exponential rise (with an e-folding time of 15 s) to a peak brightness of 37 MR (1 MR = 10^{12} photons $\text{cm}^{-2} \text{s}^{-1} 4\pi \text{sr}^{-1}$) followed by an exponential decline (with a (1/e)-folding time of 63 s). Other auroral emissions remain virtually unchanged in intensity and morphology during the flare event. The summed image of this event (Fig. 1b) shows that the flare occurs within a region of fainter, more diffuse emissions inside the main oval.

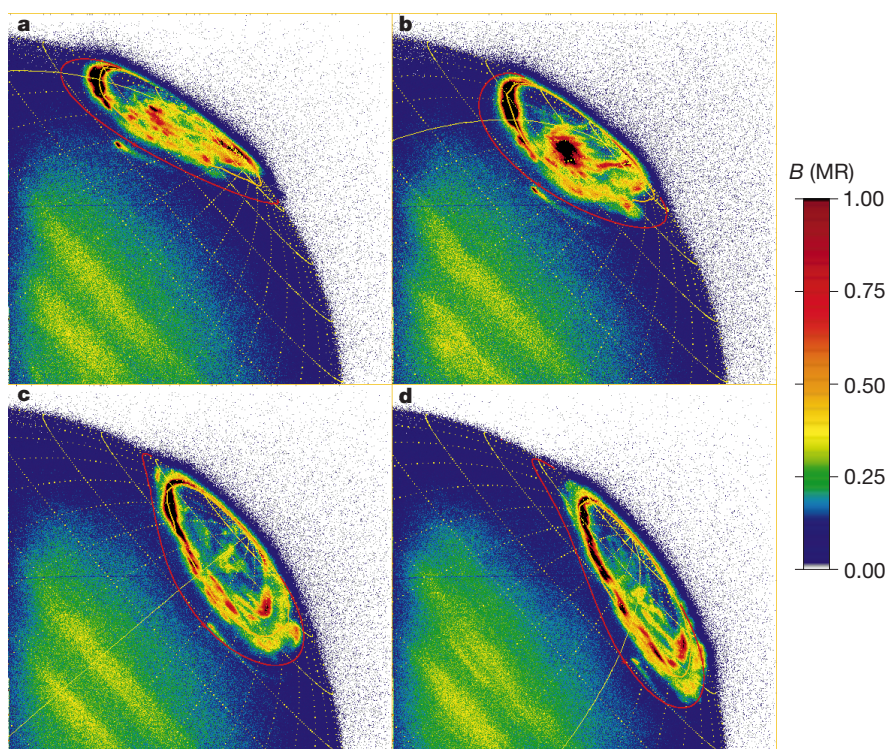


Figure 1 Hubble Space Telescope images of Jupiter's northern aurora on 21 September 1999. False-colour brightnesses are indicated in megarayleighs (MR). We have clipped the brightnesses above 1 MR to highlight fainter emissions. The exposures were made at **a**, 20:27:30–20:31:30; **b**, 21:04:50–21:08:50; **c**, 22:01:54–22:06:54; and **d**, 22:38:52–22:45:32 UT, with effective exposure times of 224.5 s, 210.5 s, 238.4 s, and 319.5 s, respectively. A jovicentric graticule with 10° intervals is overlaid (with the 180°

meridian in solid colour), along with $L = 30$ footprint and the $L = 5.9$ mapping of the Io torus. These observations were made using the F25SRF2 filter, which transmits primarily auroral emissions of the H_2 Lyman bands. The sensitivity of STIS with this filter is calculated to be 0.22 counts $\text{s}^{-1} \text{pixel}^{-1} \text{MR}^{-1}$ of H_2 and H auroral photons emitted.

A similar event, though shorter-lived and less intense than the flare, occurs in the first time-tagged observation as well (Fig. 1a), which was conducted half an hour earlier, from 20:27:30 to 20:31:30 UT. The most intense phase of this event lasts only about 10 s and reaches a maximum brightness of 17 MR. No transient brightenings are seen in the third and fourth observations (from 22:01:54 to 22:06:54 UT and from 22:38:52 to 22:45:32 UT, respectively). However, in the summed images (Fig. 1c and d), the data show fainter but distinct localized emission features near a System III longitude of 180° . The relation of these features to the localized brightenings observed earlier is unknown; however, both these fainter features and the transient brightenings occur near the noon magnetic meridian, which suggests that it may be more appropriate to consider them as being organized in magnetic local time rather than System III longitude.

The four time-tagged imaging observations were performed in short segments half an hour to an hour apart and do not present a continuous picture of the evolution of polar cap auroral activity during the two-hour observation period. However, it is clear from the first time-tagged observation that polar cap activity had begun some time before 20:27:30 UT and that similar localized brightenings had occurred at the same region as the flare. In the following discussion we focus on the flare as the most dramatic of these events.

The location of the flare at latitudes poleward of the main oval indicates that it is linked by magnetic field lines to a region of the magnetosphere that lies at radial distances beyond $30 R_J$ on Jupiter's day side. This region is not described by any of the existing jovian magnetic field models^{11,12}. We therefore used a recently developed magnetohydrodynamic (MHD) model of the solar wind interaction with the jovian magnetosphere¹³ to identify the possible source region of the charged particles responsible for the flare and thus to locate the site of the magnetospheric disturbance that triggered it.

According to the model, the flare maps to a longitudinally extended region located at joviocentric distances between 40 and $60 R_J$ in the morning sector. (Because of the corotation lag of plasma flows in the outer magnetosphere, the source region is at a somewhat earlier local time than is the flare on the planet.) It can be assumed the flare was triggered by a disturbance in this region of the magnetosphere. However, the nature and cause of such a disturbance is not known.

Unfortunately, no *in situ* data on conditions in the day side magnetosphere are available for the times of our observations, which were performed while the Galileo spacecraft was located in Jupiter's dusk magnetosphere following its fly-by of Callisto five days earlier, on 16 September. The Galileo magnetometer data show no unusual conditions at the time of the Hubble Space Telescope (HST) observations (K. Khurana, personal communication). Other Galileo particles and fields data are not available for this day.

Likewise, there are no near-Jupiter solar-wind data available for this time period. However, using a one-dimensional, time-dependent eulerian finite-difference fluid code^{14,15}, we propagated the one-hour averaged solar-wind data from the Solar Wind Electron Proton Alpha Monitor (SWEPAM)¹⁶ on the Advanced Composition Explorer (ACE) spacecraft at L1—proton speed, density, and dynamic pressure—from 1 AU to the location of Jupiter at 4.9 AU. The results (Fig. 3) show a series of sharp rises in the solar-wind dynamic pressure at the orbit of Jupiter near the time that the flare

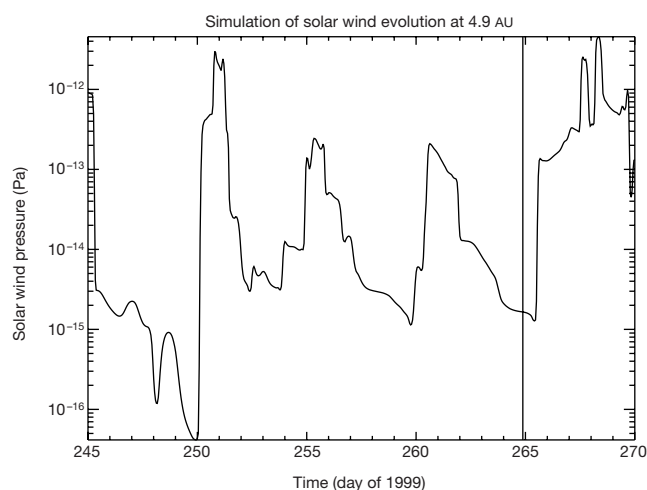


Figure 3 Modelled solar wind dynamic pressure at Jupiter. The calculated solar wind plasma pressure during days 245–270, 1999, is based on spacecraft measurements at Earth's L1 point (Earth was less than 45° from Jupiter as seen from the Sun at the time of the observations). The vertical line marks the flare event time at Jupiter.

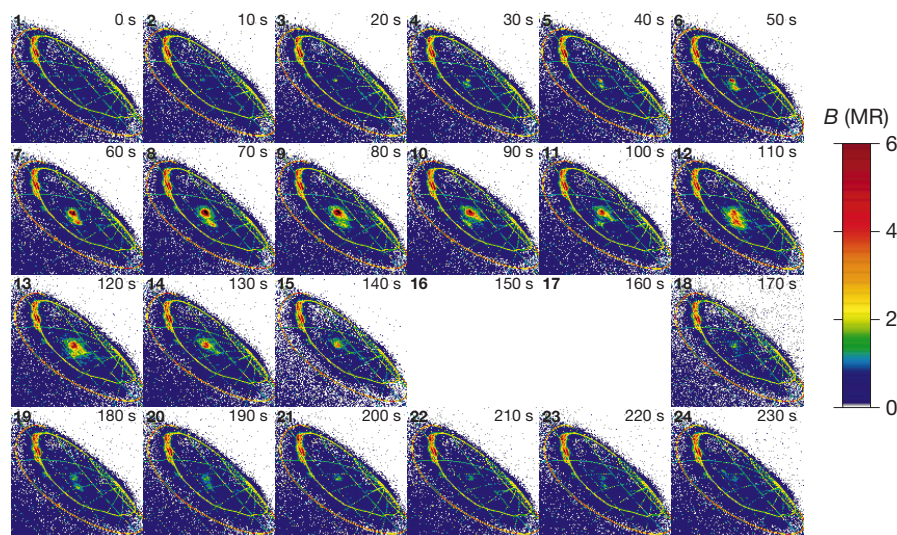


Figure 2 The rapidly evolving auroral flare observed poleward of Jupiter's auroral oval. The time series shows the evolution of the flare beginning at 21:04:50 UT on 21 September. Ten-second frames are used, and the brightness scale is as indicated (with brightnesses

clipped at 6 MR). The peak brightness of the flare at the 70-s frame is 37 MR (total $H_2 + H$ emissions). An animated version of these observations can be viewed as Supplementary Information, and at <http://pluto.space.swri.edu/yosemite/jupiter/flare.html>.

was observed. Jupiter's magnetosphere has been known since the Pioneer encounters to be highly compressible and to vary rapidly and dramatically in size with changes in solar wind dynamic pressure¹⁷. Moreover, previous studies have demonstrated a correlation between changes in solar-wind pressure and changes in the intensity of auroral emissions at both infrared and radio wavelengths^{3,4}. Although uncertainties in the simulation, and the fact that Jupiter and ACE were offset from each other by about 43° in ecliptic longitude, prevent us from identifying an exact solar-wind feature to match the flare observation, it is not unreasonable to speculate that a sharp jump in dynamic pressure at Jupiter's day side magnetopause produced the disturbance that manifested itself in the polar-cap flare. The solar-wind conditions at this time were not unusual, which suggests that such flares, if indeed triggered by changes in solar-wind pressure, may not be uncommon. We note that responses of Earth's aurora to solar-wind dynamic pressure pulses are well known, both as rapid global brightenings associated with the passage of interplanetary shocks^{18,19} and as smaller-scale transient auroral events observed in the day side cusp/cleft region²⁰. Thus similar physical processes may occasionally be at work in both the terrestrial and the jovian auroras, although the transient intensification observed at Jupiter is much more powerful than similar events observed at Earth.

A unique opportunity to test the possible role of solar-wind pressure pulses in triggering polar-cap transients was presented by the Cassini fly-by of Jupiter in December 2000. Coordinated Cassini, HST and Galileo observations were planned to provide near-simultaneous monitoring of solar-wind conditions (Cassini), auroral activity (HST), and magnetospheric conditions (Galileo). The results of these coordinated observations are expected to provide additional new examples of the puzzling new auroral feature reported here and to yield new insights into jovian auroral phenomena and the underlying magnetospheric processes. □

Received 4 October 2000; accepted 1 March 2001.

- Bhardwaj, A. & Gladstone, G. R. Auroral emissions of the giant planets. *Rev. Geophys.* **38**, 295–353 (2000).
- Hill, T. W., Dessler, A. J. & Goertz, C. K. in *Physics of the Jovian Magnetosphere* (ed. Dessler, A. J.) 353–394 (Cambridge Univ. Press, Cambridge, 1983).
- Desch, M. D. & Barrow, C. H. Direct evidence for solar wind control of Jupiter's hectometer-wavelength radio emission. *J. Geophys. Res.* **89**, 6819–6823 (1984).
- Baron, R. L., Owen, T., Connerney, J. E. P., Satoh, T. & Harrington, J. Solar wind control of Jupiter's H₃⁺ auroras. *Icarus* **120**, 437–442 (1996).
- Clarke, J. T. *et al.* Far-ultraviolet imaging of Jupiter's aurora and the Io "footprint". *Science* **274**, 404–409 (1996).
- Clarke, J. T. *et al.* Hubble Space Telescope imaging of Jupiter's UV aurora during the Galileo orbiter mission. *J. Geophys. Res.* **103**, 20217–20236 (1998).
- Prangé, R. *et al.* Detailed study of FUV Jovian auroral features with the post-COSTAR HST faint object camera. *J. Geophys. Res.* **103**, 21095–20215 (1998).
- Connerney, J. E. P., Baron, R., Satoh, T. & Owen, T. Images of excited H₃⁺ at the foot of the Io flux tube in Jupiter's atmosphere. *Science* **262**, 1035–1038 (1993).
- Satoh, T. & Connerney, J. E. P. Jupiter's H₃⁺ emissions viewed in corrected jovimagnetic coordinates. *Icarus* **141**, 236–252 (1999).
- Vasavada, A. R. *et al.* Jupiter's visible aurora and Io footprint. *J. Geophys. Res.* **104**, 27133–27142 (1999).
- Connerney, J. E. P., Acuña, M. H., Ness, N. F. & Satoh, T. New models of Jupiter's magnetic field constrained by the Io flux tube footprint. *J. Geophys. Res.* **103**, 11929–11939 (1998).
- Khurana, K. K. Euler potential models of Jupiter's magnetospheric field. *J. Geophys. Res.* **102**, 11295–11306 (1997).
- Ogino, T., Walker, R. J. & Kivelson, M. G. A global magnetohydrodynamic simulation of the jovian magnetosphere. *J. Geophys. Res.* **103**, 225–235 (1998).
- Stone, J. M. & Norman, M. L. ZEUS-2D: A radiation magnetohydrodynamics code for astrophysical flows in two space dimensions. I. The hydrodynamic algorithms and tests. *Astrophys. J. Suppl. Ser.* **80**, 753–790 (1992).
- Gosling, J. T. & Riley, P. The acceleration of slow coronal mass ejections in the high-speed solar wind. *Geophys. Res. Lett.* **23**, 2867–2870 (1996).
- McComas, D. J. *et al.* Solar wind electron proton alpha monitor (SWEPAM) for the Advanced Composition Explorer. *Space Sci. Rev.* **86**, 563–612 (1998).
- Intriligator, D. S. & Wolfe, J. H. in *Jupiter: Studies of the Interior, Atmosphere, Magnetosphere, and Satellites* (ed. Gehrels, T.) 848–869 (Univ. Arizona Press, Tucson, 1976).
- Craven, J. D., Frank, L. A., Russell, C. T., Smith, E. J. & Lepping, R. P. in *Solar Wind–Magnetosphere Coupling* (eds Kamide, Y. & Slavin, J. A.) 367–380 (Terra Scientific, Tokyo, 1986).
- Zhou, X. & Tsurutani, B. T. Rapid intensification and propagation of the dayside aurora: Large scale interplanetary pressure pulses (fast shocks). *Geophys. Res. Lett.* **26**, 1097–1100 (1999).

- Sandholt, P. E. *et al.* Cusp/cleft auroral activity in relation to solar wind dynamic pressure, interplanetary magnetic field B_z and B_y. *J. Geophys. Res.* **99**, 17323–17342 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>).

Acknowledgements

We gratefully acknowledge discussions with B. T. Tsurutani and J. L. Burch and the support of the Space Telescope Science Institute.

Correspondence and requests for materials should be addressed to J.H.W. (e-mail: hunterw@umich.edu).

Quantum computing in molecular magnets

Michael N. Leuenberger & Daniel Loss

Department of Physics and Astronomy, University of Basel, Klingelbergstrasse 82, 4056 Basel, Switzerland

Shor and Grover demonstrated that a quantum computer can outperform any classical computer in factoring numbers¹ and in searching a database² by exploiting the parallelism of quantum mechanics. Whereas Shor's algorithm requires both superposition and entanglement of a many-particle system³, the superposition of single-particle quantum states is sufficient for Grover's algorithm⁴. Recently, the latter has been successfully implemented⁵ using Rydberg atoms. Here we propose an implementation of Grover's algorithm that uses molecular magnets^{6–10}, which are solid-state systems with a large spin; their spin eigenstates make them natural candidates for single-particle systems. We show theoretically that molecular magnets can be used to build dense and efficient memory devices based on the Grover algorithm. In particular, one single crystal can serve as a storage unit of a dynamic random access memory device. Fast electron spin resonance pulses can be used to decode and read out stored numbers of up to 10⁵, with access times as short as 10^{–10} seconds. We show that our proposal should be feasible using the molecular magnets Fe₈ and Mn₁₂.

Suppose we want to find a phone number in a phone book consisting of $N = 2^n - 1$ entries. Usually it takes $N/2$ queries on average to be successful. Even if the N entries were encoded in binary, a classical computer would need approximately $\log_2 N$ queries to find the desired phone number². But the computational parallelism provided by the superposition and interference of quantum states enables the Grover algorithm to reduce the search to one single query². Here we will show that this query can be implemented in terms of a unitary transformation applied to the single spin of a molecular magnet. Such molecular magnets, forming identical and largely independent units, are embedded in a single crystal so that the ensemble nature of such a crystal provides a natural amplification of the magnetic moment of a single spin. However, for the Grover algorithm to succeed, it is necessary to find ways to generate arbitrary superpositions of spin eigenstates. For spins larger than 1/2 this turns out to be a highly non-trivial task as spin excitations induced by magnetic dipole transitions in conventional electron spin resonance (ESR) can occur only in discrete steps of one $\hbar$ (Planck's constant divided by 2π), that is, single steps by two or more $\hbar$ values are excluded by selection rules. To circumvent such physical limitations we propose an unusual scenario which, in principle, allows the controlled generation of arbitrary spin superpositions through the use of multifrequency coherent magnetic

radiation in the microwave and radiofrequency range. In particular, we will show by means of the S -matrix and time-dependent high-order perturbation theory that by using advanced ESR techniques it is possible to coherently populate and manipulate many spin states simultaneously by applying one single pulse of a magnetic a.c. field containing an appropriate number of matched frequencies. This a.c. field creates a nonlinear response of the magnet through multiphoton absorption processes involving particular sequences of σ and π photons which allows the encoding and, similarly, the decoding of states. Finally, the subsequent read-out of the decoded quantum state can be achieved by means of pulsed ESR techniques. These exploit the non-equidistance of energy levels which is typical of molecular magnets. The method we propose here is interesting in its own right since there has never been an experimental or theoretical attempt, to our knowledge, that shows that the states of spin systems with $s > 1/2$ can be coherently populated.

We implement the Grover algorithm in a single-spin representation with the level spectrum of a spin system as shown in Fig. 1. First, a strong magnetic field in the z direction must be applied to prepare the initial state $|\psi_0\rangle = |s\rangle$. Then this field is reduced almost to zero (up to the bias δH_z) in such a way that all $|m\rangle$ -states are localized, say, on the left side of the potential barrier. Thus, the magnetic moment pointing along the z axis assumes its maximum value and the single spin of a molecular magnet is described by the hamiltonian $H_{\text{spin}} = H_a + V$ (see Fig. 1). To mark specific states, for example, with a certain occupation amplitude (including phases), we could apply weak oscillating transverse magnetic fields $\mathbf{H}_\perp$ to induce multiphoton transitions via virtual states, which can usually be calculated in perturbation theory in $\mathbf{H}_\perp$. However, the Grover algorithm requires that all the k -photon transitions, $k = 1, 2, \dots, s-1$, have (approximately) the same amplitudes (and possibly different phases). Thus, this would require that all terms of power $V^1, V^2, \dots, V^{s-1}$ must be of comparable magnitude. Obviously, perturbation theory breaks down in such a case.

To bypass this problem, in our method, all the transition amplitudes between the states $|s\rangle$ and $|m\rangle$, $m = 1, 2, \dots, s-1$, are of the same order in perturbation V . This allows us to use perturbation theory. It works only if the energy levels are not equidistant, which is typically the case in molecular magnets owing to anisotropies (in contrast to, for example, a harmonic oscillator potential). In general, if we choose to work with the states $m = m_0, m_0 + 1, \dots, s-1$, where $m_0 = 1, 2, \dots, s-1$, we have to go up to the n th order in perturbation, where $n = s - m_0$ and $n-1$ is the number of computational states used for the Grover search algorithm (see below), to obtain the first non-vanishing contribution. Figure 2 shows the transitions for $s = 10$ and $m_0 = 5$. The n th-order transitions correspond to the nonlinear response of the spin system to strong magnetic fields. Thus, a coherent magnetic pulse of duration T is needed with a discrete frequency spectrum $\{\omega_m\}$, say, for Mn_{12} between 20 and 120 GHz and a single low-frequency ω_0 around 100 MHz (for pulse-shaping techniques see ref. 11 and references therein). The low-frequency field $\mathbf{H}_z(t) = H_0(t)\cos(\omega_0 t)\mathbf{e}_z$, applied along the easy axis, couples to the spin of the molecular magnet through the hamiltonian

$$V_{\text{low}}(t) = g\mu_B H_0(t)\cos(\omega_0 t)S_z \quad (1)$$

where $\hbar\omega_0 \ll \epsilon_{m_0} - \epsilon_{m_0+1}$ and $\mathbf{e}_z$ is the unit vector pointing along the z axis. The π photons¹² of V_{low} supply the necessary energy for the resonance condition (see below). They give rise to virtual transitions with $\Delta m = 0$, that is, they do not transfer any angular momentum, see Fig. 2.

The perturbation hamiltonian for the high-frequency transitions from $|s\rangle$ to virtual states that are just below $|m\rangle$, $m = m_0, \dots, s-1$, given by the transverse fields $\mathbf{H}_\perp(t) =$

$\Sigma_{m=m_0}^{s-1} H_m(t)[\cos(\omega_m t + \Phi_m)\mathbf{e}_x - \sin(\omega_m t + \Phi_m)\mathbf{e}_y]$, reads

$$V_{\text{high}}(t) = \sum_{m=m_0}^{s-1} g\mu_B H_m(t)[\cos(\omega_m t + \Phi_m)S_x - \sin(\omega_m t + \Phi_m)S_y] \quad (2)$$

$$= \sum_{m=m_0}^{s-1} \frac{g\mu_B H_m(t)}{2} [e^{i(\omega_m t + \Phi_m)} S_+ + e^{-i(\omega_m t + \Phi_m)} S_-]$$

with phases Φ_m (see below), where we have introduced the unit vectors $\mathbf{e}_x$ and $\mathbf{e}_y$ pointing along the x and y axes, respectively. These transverse fields rotate clockwise and thus produce left circularly polarized σ^- photons which induce only transitions in the left well (see Fig. 1). In general, absorption (emission) of σ^- photons gives rise to $\Delta m = -1$ ($\Delta m = +1$) transitions, and vice versa in the case of σ^+ photons. Anti-clockwise rotating magnetic fields of the form $\mathbf{H}_\perp(t) = \Sigma_{m=m_0}^{s-1} H_m(t)[\cos(\omega_m t + \Phi_m)\mathbf{e}_x - \sin(\omega_m t + \Phi_m)\mathbf{e}_y]$ can be used to induce spin transitions only in the right well (see Fig. 1). In this way, both wells can be accessed independently.

Next we calculate the quantum amplitudes for the transitions induced by the magnetic a.c. fields (see Fig. 2) by evaluating the S -matrix perturbatively. The j th-order term of the perturbation series of the S -matrix in powers of the total perturbation hamiltonian $V(t) = V_{\text{low}}(t) + V_{\text{high}}(t)$ is expressed by

$$S_{m,s}^{(j)} = \left(\frac{1}{i\hbar}\right)^j \prod_{k=1}^{j-1} \int_{-\infty}^{\infty} dt_k \times \int_{-\infty}^{\infty} dt_j \Theta(t_k - t_{k+1}) U(\infty, t_1) V(t_1) U(t_1, t_2) \times V(t_2) \dots V(t_j) U(t_j, -\infty) \quad (3)$$

which corresponds to the sum over all Feynman diagrams of order j , and where $U(t, t_0) = e^{-iH(t-t_0)/\hbar}$ is the free propagator, $\Theta(t)$ is the Heavyside function. The total S -matrix is then given by $S = \Sigma_{j=0}^{\infty} S^{(j)}$. The high-frequency virtual transition changing m from s to $s-1$ is induced by the frequency $\omega_{s-1} = \omega_{s-1,s} - (n-1)\omega_0$. The other high frequencies ω_m , $m = m_0, \dots, s-2$, of the high-frequency fields H_m mismatch the level separations by ω_0 , that is, $\hbar\omega_m = \epsilon_m - \epsilon_{m+1} + \hbar\omega_0$, see Fig. 2. As the levels are not equidistant, it is possible to choose the low and high frequencies in such a way that $S_{m,s}^{(j)} = 0$ for $j < n$, in which case the resonance condition is not satisfied, that is, energy is not conserved. In addition, the higher-order amplitudes $|S_{m,s}^{(j)}|$ are negligible compared to $|S_{m,s}^{(n)}|$ for $j > n$. Using rectangular pulse shapes with $H_k(t) = H_k$, if $-T/2 < t < T/2$, and 0 otherwise, for $k = 0$ and $k \geq m_0$, we obtain after lengthy but straightforward calculation ($m \geq m_0$)

$$S_{m,s}^{(n)} = \sum_F \Omega_m \frac{2\pi}{i} \left(\frac{g\mu_B}{2\hbar}\right)^n \frac{\Pi_{k=m}^{s-1} H_k e^{i\Phi_k} H_0^{m-m_0} p_{m,s}(F)}{(-1)^{q_F} q_F! r_s(F)! \omega_0^{n-1}} \times \delta^{(T)}\left(\omega_{m,s} - \sum_{k=m}^{s-1} \omega_k - (m-m_0)\omega_0\right) \quad (4)$$

where $\Omega_m = (m-m_0)!$ is the symmetry factor of the Feynman diagrams F (see Fig. 2), $q_F = m - m_0 - r_s(F)$, $p_{m,s}(F) = \Pi_{k=m}^{s-1} \langle k|S_z|k\rangle^{r_k(F)} \Pi_{k=m}^{s-1} \langle k|S_-|k+1\rangle$, where $r_k(F) = 0, 1, 2, \dots \leq m - m_0$ is the number of π transitions directly above or below the state $|k\rangle$, depending on the particular Feynman diagram F , and $\delta^{(T)}(\omega) = 1/2\pi \int_{-T/2}^{T/2} e^{i\omega t} dt = \sin(\omega T/2)/\omega$ is the delta-function of width $1/T$, ensuring overall energy conservation (the resonance condition) for $\omega T \gg 1$. The duration T of the magnetic pulses must be shorter than the lifetimes τ_d of the states $|m\rangle$ (see Fig. 1). For illustration, we now focus on the case $s = 10$, $m_0 = 5$, and thus $n = 5$, described by $S_{m,10}^{(5)}$, since this is most relevant for the molecular magnets Mn_{12} and Fe_8 (see Fig. 2). In general, the Grover algorithm requires that the levels are simultaneously populated with roughly equal amplitudes, that is, $|S_{m,s}^{(n)}| \approx |S_{s-1,s}^{(n)}|$, for all $m \geq m_0$,

from which we can deduce the required field amplitudes using equation (4)

$$\begin{aligned} H_8/H_0 &= -0.04, \quad H_7/H_0 = -0.25, \quad H_6/H_0 = -0.61, \\ H_5/H_0 &= -1.12 \end{aligned} \quad (5)$$

This means that the fields H_0 and H_9 , and the frequency ω_0 can be chosen independently. We note particularly that the amplitudes H_k do not differ too much from each other, which can be traced back to partial cancellations in equation (4) owing to the factor $(-1)^{q_i}$. This fact is most useful for applications.

Next, we estimate the transition rate needed to coherently populate the five levels with one single pulse with $V_{\text{low}} + V_{\text{high}}$. As $[\delta^{(T)}(\omega)]^2 \approx (T/2)\pi\delta^{(T)}(\omega)$, the transition rate $w_{m,s} = |S_{m,s}^{(n)}|^2/T$ from $|10\rangle$ to $|5\rangle$ at resonance (with $\delta^{(T)}(0) = T/2\pi$) reads

$$w_{5,10} = T \left(\frac{g\mu_B}{2\hbar} \right)^{10} \left| \frac{H_5 H_6 H_7 H_8 H_9 p_{5,10}}{4! \omega_0^4} \right|^2 \quad (6)$$

where $p_{5,10} = \prod_{k=5}^9 \langle k|S_-|k+1\rangle$. We then insert the parameters $\omega_0 = 5 \times 10^8 \text{ s}^{-1}$, $T = 10^{-7} \text{ s}$, $H_0 = H_9 = 20 \text{ G}$ giving the transition rate $w_{5,10} = 9 \times 10^6 \text{ s}^{-1}$, which is of the order of τ_d , that is, $T w_{5,10} = 1$. For our purpose it is sufficient to choose $T w_{5,10} \ll 1$. If $\omega_0 = 5 \times 10^7 \text{ s}^{-1}$, $T = 10^{-7} \text{ s}$ and $H_0 = H_9 = 2 \text{ G}$, we obtain the transition rate $w_{5,10} = 9 \times 10^4 \text{ s}^{-1}$, and thus $T w_{5,10} \approx 0.01$. Thus we have shown that the required amplitudes and frequencies of the fields given in equation (1) and (2) are experimentally accessible.

We adapt now the Grover scheme⁵ to describe the quantum computational read-in and decoding of the quantum data register $\mathbf{a}$. For simplicity we set the relative phase $\Phi_0 = 0$ of the low-frequency field $\mathbf{H}_z(t)$; see equation (1).

(1) Read-in. We start from the ground state $|s\rangle$ as initial state. Then, in order to introduce the desired phases Φ_m for each state $|m\rangle$ we need to irradiate the system with a coherent magnetic pulse of duration T containing the $n-1$ high-frequency fields $H_m [\cos(\omega_m t + \Phi_m) \mathbf{e}_x - \sin(\omega_m t + \Phi_m) \mathbf{e}_y]$ (see equation (2)) with $\Phi_m = \sum_{k=s-1}^{m+1} \Phi_k + \varphi_m$ (φ_m are the relative phases), and the low-

frequency field $\mathbf{H}_z(t)$ (see equation (1)), yielding $S_{m_0,s}^{(n)} = \dots = S_{s-2,s}^{(n)} = S_{s-1,s}^{(n)} = \pm \eta$, $\eta > 0$, that is, $\varphi_m = 0, \pi$, depending on the number to be encoded. For example, to encode the number $13_{10} = 1101_2$ in Fig. 2, we need $\varphi_9 = \varphi_8 = \varphi_7 = 0$ and $\varphi_6 = \varphi_5 = \pi$, where the states $m = 9, 8, 7, 6, 5$ represent the binary digits $2^0, 2^1, 2^2, 2^3, 2^4$, respectively. We note that $S_{m,s}^{(n)}$ can be either positive or negative, depending on the explicit Feynman diagrams F . In this way we can prepare the quantum data register $\mathbf{a} = (a_s, a_{s-1}, \dots, a_{m_0})$, where the bits $a_i = 1$, $a_m = \pm \eta$ are the amplitudes of the state $|\psi\rangle = \sum_{m=m_0}^s a_m |m\rangle$. This pulse performs a unitary transformation up to order η , that is, $U_{m,s} + O(\eta^2) = S_{m,s}^{(n)}$. In this way an arbitrary integer between 0 and 2^{n-1} can be stored.

(2) Decoding. In order to decode the phase information stored in the data register $\mathbf{a}$, a universal single pulse (see equations (1) and (2)) leading to $S_{m_0,s}^{(n)} = \dots = S_{s-2,s}^{(n)} = S_{s-1,s}^{(n)} = -\eta$ must be applied, which performs approximately a unitary transformation for $\eta \ll 1$. The accumulated error is about $n\eta^2$, which must be kept smaller than 1. Because $\eta > \eta_0 > 0$, with η_0 given by the precision of the detection, we require that $n \ll 1/\eta_0^2$. We note that this decoding works also if the bits a_m , with $\eta_0 < |a_m| \ll 1$, have different amplitudes (but still of similar magnitude). This pulse amplifies the flipped bits, which have amplitude $-\eta$, and suppresses the rest of the bits with amplitude $+\eta$. For the example shown in Fig. 2 the relative phases in equation (2) must be set $\varphi_9 = \varphi_7 = \varphi_5 = 0$ and $\varphi_8 = \varphi_6 = \pi$ (irrespective of $\mathbf{a}$).

(3) Read-out. Once a state has been marked and amplified, that is, decoded, we must be able to read out this information physically. This task can be accomplished by standard spectroscopy with, say, pulsed ESR, where the circularly polarized radiation can now be incoherent because we need the absorption intensity of only one pulse. Full spectral analysis of Mn_{12} has been performed with ESR¹³ and neutron scattering¹⁴. Thus, we can assume that the spectrum is known. Now, irradiation of the magnet with a single pulse of duration T containing the frequencies $\omega_{m-1,m}$, where $m = s-2, \dots, m_0$, induces transitions that are described by the first-order amplitudes $S_{m-1,m}^{(1)}$ (higher-order multiphoton effects

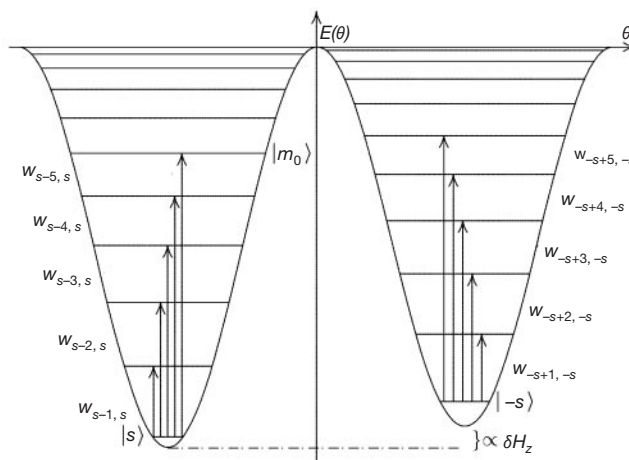


Figure 1 Double well potential seen by the spin due to magnetic anisotropies in Mn_{12} . Arrows depict transitions between spin eigenstates driven by the external magnetic field $\mathbf{H}$ (see below). Molecular magnets have the important advantage that they can be grown naturally as single crystals of up to 10–100 μm in length containing about 10^{12} to 10^{15} (largely) independent units so that only minimal sample preparation is required. The molecular magnets behave like single spins: for example, Mn_{12} (refs 7, 8) and Fe_8 (refs 9, 10) have a spin $s = 10$ ground state. Thus, they can be described by a single-spin hamiltonian of the form $H_{\text{spin}} = H_a + V + H_{\text{sp}} + H_1^{15,16}$, where $H_a = -AS_z^2 - BS_z^4$ represents the magnetic anisotropy ($A \gg B > 0$), that is, the easy axis of the spin lies along the z direction. The Zeeman term $V = g\mu_B \mathbf{H} \cdot \mathbf{S}$ describes the coupling between the external magnetic field $\mathbf{H}$ and the spin $\mathbf{S}$ of length s . The calculational states are given by

the $2s+1$ eigenstates $|m\rangle$ of $H_a + g\mu_B \delta H_z S_z$ with eigenenergies $\epsilon_m = -Am^2 - Bm^4 + g\mu_B \delta H_z m$, $-s \leq m \leq s$. The corresponding classical anisotropy potential energy $E(\theta) = -As \cos^2 \theta - Bs \cos^4 \theta + g\mu_B \delta H_z s \cos \theta$, shown here, is obtained by the substitution $S_z = s \cos \theta$, where θ is the polar spherical angle. We have introduced the notation $\hbar \omega_{m,m'} = \epsilon_m - \epsilon_{m'}$. The hamiltonian H_1 induces tunnelling between (quasi) degenerate $|m\rangle$ -states with tunnel splitting energy $E_{mm'}$. However, applying a bias field δH_z such that $g\mu_B \delta H_z > E_{mm'}$, tunnelling can be completely suppressed and thus H_1 can be neglected^{15,16}. We also assume temperatures of below 1 K such that transitions due to spin–phonon interactions (H_{sp}) can also be neglected. In this regime, the level lifetime in Fe_8 and Mn_{12} is estimated to be about $\tau_d = 10^{-7} \text{ s}$, limited mainly by hyperfine and/or dipolar interactions^{10,15,16}.

can now be neglected for sufficiently small fields). For instance, let us assume that the state $|7\rangle$ is marked, that is, populated. Then, we would observe stimulated emission for the transitions from $|7\rangle$ to $|8\rangle$ at the frequency $\omega = \omega_{7,8}$ and stimulated absorption of approximately the same intensity for the transition from $|7\rangle$ to $|6\rangle$ at $\omega = \omega_{6,7}$, which uniquely identifies the marked level, because the levels are not equidistant. Generally, if the states $|m_1\rangle, |m_2\rangle, \dots, |m_k\rangle$, where $1 < k < n$, are marked, the following absorption/emission intensity in leading order is measured:

$$I_{s-2}^{m_0} = \sum_{i=1}^k (|S_{m_i-1, m_i}^{(1)}|^2 + |S_{m_i+1, m_i}^{(1)}|^2) \quad (7)$$

This spectrum identifies all the marked states unambiguously. We

emphasize that the entire Grover algorithm (read-in, decoding, read-out) requires three subsequent pulses each of duration T with $\tau_d > T > \omega_0^{-1} > \omega_m^{-1}, \omega_{m,m\pm 1}^{-1}$. This gives a 'clock-speed' of about 10 GHz for Mn_{12} , that is, the entire process of read-in, decoding, and read-out can be performed within about 10^{-10} s.

Finally, so far we have used only the left well of the potential in Fig. 1. As both wells can be accessed separately during read-in, decoding and read-out by magnetic fields that rotate either clockwise or anti-clockwise, a single molecular magnet represents a two-digit number. Every digit can store $N = 2^{s-1}$ elements at most, which allows us to store a number between 0 and $2^{2s-2} = 2.6 \times 10^5$ in a single crystal made of molecular magnets with spin $s = 10$. If $M > 2$ phases can be distinguished for φ_m (depending on the

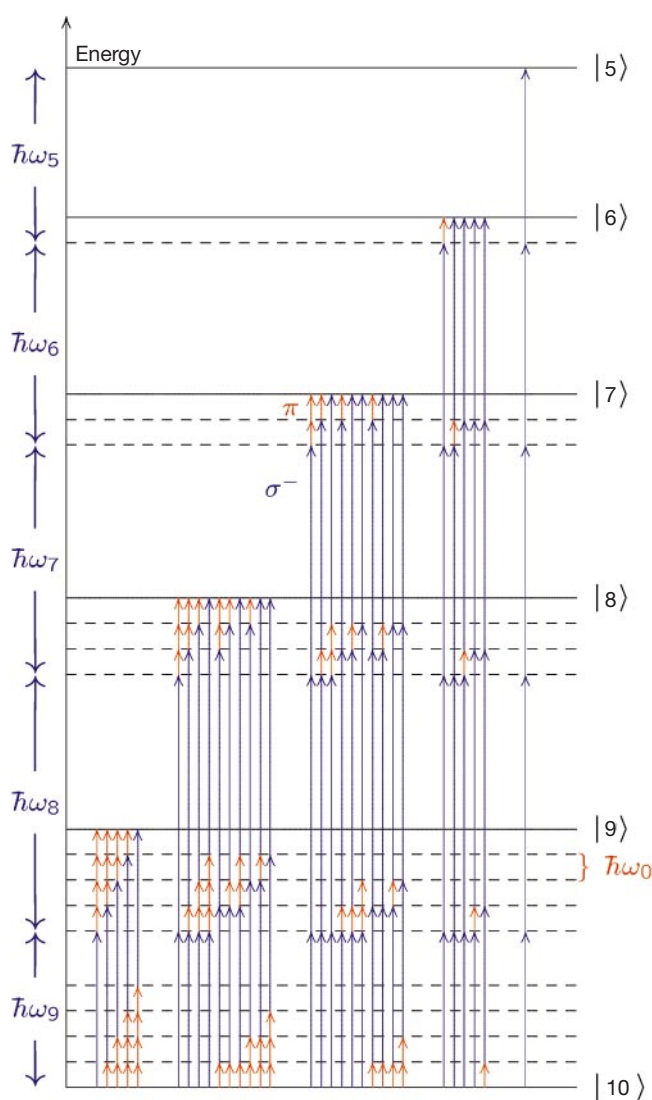


Figure 2 Feynman diagrams F that contribute to $S_{m,s}^{(5)}$ for $s = 10$ and $m_0 = 5$ describing transitions (of 5th order in V) in the left well of the spin system (see Fig. 1). The blue (red) arrows indicate the transitions induced by the high-(low)-frequency magnetic fields H_m (H_0) shown in equation (2) (equation (1)). Whereas the H_m fields transfer angular momentum to the spin of the molecular magnet by means of σ^- photons, the H_0 field provides only energy without angular momentum by means of π photons. In this way all the transition amplitudes $S_{m,s}^{(j)}$ are of similar magnitude (see text). We note that $S_{m,s}^{(j)} = 0$ for $j < n$, and $|S_{m,s}^{(j)}| \ll |S_{m,s}^{(n)}|$ for $j > n$. For example, the transition from $|10\rangle$ to $|7\rangle$ arises from the absorption of five photons in total, comprising three σ^- photons and two π photons. As it does not matter in which order the five photons are absorbed, there are $\binom{5}{2} = 10$ different kinds of diagrams. Also, it is essential that the levels are not

equidistant; if they are, this scheme for multiphoton absorption does not work, because then resonances can already occur in lower orders of the S -matrix leading to incoherent populations of the levels. There is a global phase factor due to the unperturbed time evolution of the spin system, given by $e^{-i\epsilon_m(t-t_0)/\hbar}$, which can easily be accounted for and shall thus be ignored here. For a first test of the nonlinear response described here, we can irradiate the molecular magnet with an a.c. field of frequency $\omega_{s-1,s}/2$, which gives rise to a two-photon absorption and thus to a Rabi oscillation between the states $|s\rangle$ and $|s-1\rangle$. We note that for stronger magnetic fields it is in principle possible to generate superpositions of Rabi oscillations between the states $|s\rangle$ and $|s-1\rangle$, $|s\rangle$ and $|s-2\rangle$, $|s\rangle$ and $|s-3\rangle$, and so on (to be published elsewhere).

experimental resolution)⁵, a number between 0 and M^{2s-2} can be stored in a single crystal of molecular magnets with spin s . We note that the experimental overhead required by the Grover search algorithm involves only the control of $\log_M N$ frequencies, which, once available, can decode any number between 1 and N by means of a single magnetic pulse. Our proposal for implementing Grover's algorithm works not only for molecular magnets but for any electron or nuclear spin system with non-equidistant energy levels. Although such spin systems cannot be scaled to arbitrarily large spin s —the larger a spin becomes, the faster it decoheres and the more classical its behaviour will be—we can use such spin systems of given s to great advantage in building dense and highly efficient memory devices. □

Received 24 November 2000; accepted 14 February 2001.

- Shor, P. In *Proc. 35th Ann. Symp. Foundations of Computer Science* (ed. Goldwasser, S.) 124–134 (IEEE Computer Society Press, Los Alamitos, 1994).
- Grover, L. K. Quantum computers can search arbitrarily large databases by a single query. *Phys. Rev. Lett.* **79**, 4709–4712 (1997).
- Lloyd, S. Quantum search without entanglement. *Phys. Rev. A* **61**, R010301-1–010301-4 (1999).
- Ekert, A. K. & Jozsa, R. Quantum computation and Shor's factoring algorithm. *Rev. Mod. Phys.* **68**, 733–753 (1996).
- Ahn, J., Weinacht, T. C. & Bucksbaum, P. H. Information storage and retrieval through quantum phase. *Science* **287**, 463–465 (2000).
- Thiaville, A. & Miltat, J. Magnetism: small is beautiful. *Science* **284**, 1939–1940 (1999).
- Thomas, L. *et al.* Macroscopic quantum tunnelling of magnetization in a single crystal of nanomagnets. *Nature* **383**, 145–147 (1996).
- Friedman, J. R., Sarachik, M. P., Tejada, J. & Ziolo, R. Macroscopic measurement of resonant magnetization tunneling in high-spin molecules. *Phys. Rev. Lett.* **76**, 3830–3833 (1996).
- Sangregorio, C., Ohm, T., Paulsen, C., Sessoli, R. & Gatteschi, D. Quantum tunneling of the magnetization in an iron cluster nanomagnet. *Phys. Rev. Lett.* **78**, 4645–4648 (1997).
- Wernsdorfer, W., Sessoli, R., Caneschi, A., Gatteschi, D. & Cornia, A. Nonadiabatic Landau-Zener tunneling in Fe_8 molecular nanomagnets. *Europhys. Lett.* **50**, 552–558 (2000).
- Fitzgerald, R. Pulse shaping improves efficiency of soft X-ray harmonic generation. *Phys. Today* **53**, 24–28 (2000).
- Cohen-Tannoudji, C., Diu, B. & Laloë, F. *Quantum Mechanics* Vol. 2, 1323–1339 (Wiley, New York).
- Barra, A. L., Gatteschi, D. & Sessoli, R. High-frequency EPR spectra of a molecular nanomagnet: Understanding quantum tunneling of the magnetization. *Phys. Rev. B* **56**, 8192–8198 (1997).
- Mirebeau, I. *et al.* Low-energy magnetic excitations of the Mn_{12} -acetate spin cluster observed by neutron scattering. *Phys. Rev. Lett.* **83**, 628–631 (1999).
- Leuenberger, M. N. & Loss, D. Spin tunneling and phonon-assisted relaxation in Mn_{12} -acetate. *Phys. Rev. B* **61**, 1286–1302 (2000).
- Leuenberger, M. N. & Loss, D. Incoherent Zener tunneling and its application to molecular magnets. *Phys. Rev. B* **61**, 12200–12203 (2000).

Acknowledgements

We thank G. Salis and J. Schliemann for useful comments. This work has been supported in part by the Swiss NSF and by the European Union Molnanomag network.

Correspondence should be addressed to D.L. (e-mail: Daniel.Loss@unibas.ch).

Correlated electrons in δ -plutonium within a dynamical mean-field picture

S. Y. Savrasov, G. Kotliar & E. Abrahams

Department of Physics and Astronomy and Center for Materials Theory,
Rutgers University, Piscataway, New Jersey 08854-8019, USA

Given the practical importance of metallic plutonium, there is considerable interest^{1–3} in understanding its fundamental properties. Plutonium undergoes a 25 per cent increase in volume⁴ when transformed from its α -phase (which is stable below 400 K) to the δ -phase (stable at around 600 K), an effect that is crucial for issues of long-term storage and disposal. It has long been suspected that this unique property is a consequence of the special location of plutonium in the periodic table, on the border between the light and heavy actinides—here, electron wave-particle duality (or itinerant versus localized behaviour) is important⁵. This situation

has resisted previous theoretical treatment. Here we report an electronic structure method, based on dynamical mean-field theory, that enables interpolation between the band-like and atomic-like behaviour of the electron. Our approach enables us to study the phase diagram of plutonium, by providing access to the energetics and one-electron spectra of strongly correlated systems. We explain the origin of the volume expansion between the α - and δ -phases, predict the existence of a strong quasiparticle peak near the Fermi level and give a new viewpoint on the physics of plutonium, in which the α - and δ -phases are on opposite sides of the interaction-driven localization–delocalization transition.

Here we argue that the substantial volume expansion that occurs when metallic plutonium changes from the α - to the δ -phase is the manifestation of the competition between localization (caused by the electron–electron interactions) and delocalization (caused by the kinetic energy) tendencies. The modern theoretical tool to compute ground-state properties such as the volume of the solid is density-functional theory (DFT) in its local density or generalized gradient approximations (LDA or GGA; for a review see ref. 6). However, it fails here with unprecedented errors: the theoretical volume of the δ -phase is in error by 30%, and magnetic long-range order not observed experimentally is predicted^{7–9}. This illustrates the inability of traditional electronic structure methods to predict properties of strongly correlated systems.

The many-body theory of strongly interacting electrons on a lattice is an alternative approach with a long history. Hubbard¹⁰ was one of the first to point out the importance, in the solid state, of Coulomb correlations which occur inside atoms. Subsequently, attempts to build theories which interpolate between the atom and the solid were made. The many-body crystal wavefunction has to reduce to many-body atomic wavefunctions as lattice spacing is increased. Unfortunately, this limiting behaviour is missed in the density-functional method.

Over the past ten years, a modern many-body technique which is able to interpolate between band-like and atomic-like behaviour of electrons, the dynamical mean-field theory (DMFT), has been developed (for a review see ref. 11). It has resulted in a thorough understanding of the Mott transition in hamiltonians with strong repulsive interactions. It is of interest to consider if such a theory could be implemented for realistic multiband solids with a variety of details brought about by the crystallographic environment and the constituents involved—and if it would have the predictive power to explain such problems as the structural phase diagrams of plutonium, its alloys and compounds. The development of this electronic-structure method as an alternative to DFT would be an important advance in the physics of d - and f -electron systems in general, and in the problem of nuclear materials in particular. It is therefore a pressing problem in the area of strongly correlated electron physics^{11–13}.

Here we discuss an implementation of DMFT which allows us to incorporate the basic electronic structure of real materials into the DMFT equations. This permits the calculation of the total energy of a given compound. We show that this method leads us to the explanation of the basic features of the phase diagram of plutonium.

How does this method include atomic physics? We illustrate this point by looking at the atomic shell of an f -electron system—such as plutonium—embedded in the solid-state environment. First, we determine the spectrum of excitations for the shell, that is, a set of many-body levels describing processes of removing and adding electrons. In the simplified case, when every f electron has roughly the same kinetic energy ϵ_f and Coulomb repulsion energy U , the total energy of the shell with n electrons is given by $E_n = \epsilon_f n + U n(n-1)/2$ and the excitation spectrum is given by $\epsilon_n = E_{n+1} - E_n = \epsilon_f + U n$. The constant U can be computed from first principles. Second, we take into account that in the solid, these many-body atomic states hybridize with each other and with the

itinerant s , p and d electrons. Within the DMFT method, this process is described by a frequency-dependent hybridization function, $\Delta(\omega)$, which is a measure of the ability of the f electron to delocalize. This quantity is unknown at the outset, and needs to be determined self-consistently. The entire problem is thus reduced to solving the 'Anderson impurity model'¹⁴ (AIM) for the given f shell. This gives us the impurity Green's function $G_f(\omega)$ and self-energy $\Sigma(\omega)$ for f electrons.

We now consider how the itinerant band aspects of the f electrons are captured by this method. The bands are described by the eigenvalues of the matrix one-electron hamiltonian $H(k)$. In our implementation, $H(k)$ is obtained by constructing the LDA Kohn–Sham hamiltonian. From this, we subtract the average of the Coulomb energy of the f electrons as the latter is treated explicitly later, using a tight-binding, linear muffin-tin orbital (LMTO) basis¹⁵. A significant development in the present method is its ability to handle realistic situations in which several bands are involved.

The important step brought by the dynamical mean-field description is to require that the impurity Green's function $G_f(\omega)$ coincides with the full crystal Green's function G_c evaluated at the impurity site (site-diagonal G_c). In practice, we begin with a guess for the hybridization function, compute the impurity Green's function and then, using $H(k)$, the site-diagonal G_c . This is set equal to a new impurity $G_f(\omega)$, from which a new $\Delta(\omega)$ is determined. This iterative procedure is repeated until self-consistency in $\Delta(\omega)$ is achieved.

After the crystal Green's function is obtained, an updated charge density is computed. The latter modifies the one-electron hamiltonian $H(k)$, which in turn modifies the self-consistent solution described above. This brings us to a second self-consistent loop, the loop over charge density similar to the one in DFT. Incorporating such an external loop into the algorithm allows us to compute from first principles a central quantity of the material: its ground-state total energy. As a consequence, a phase diagram of the solid of interest can be studied by this method. By varying lattice spacing, we can look for the total-energy minimum, and predict different crystallographic configurations.

We gain several advantages by using the many-body, instead of the density-functional, description of solids. First, our approach is designed to reproduce both ground-state properties and excitation spectra. A main shortcoming of DFT is thus overcome. Second, the method gives direct access to finite temperatures via the use of temperature Green's functions. Then in principle, phase transition temperatures, both magnetic and crystallographic, should be accessible. Last, the method works naturally in the local-moment regime—when on a short timescale, magnetic moments exist on atoms, but on long timescales, there is no net moment on the average.

We now describe some details of the method that solves the AIM for given hybridization $\Delta(\omega)$ and resulting self-energy $\Sigma(\omega)$. The latter is in general a 14×14 matrix for f electrons. For the relativistic f level in cubic symmetry, it is reduced to 5×5 with 4 non-zero off-diagonal elements. Because the general AIM solution is a formidable numerical problem, we make several simplifications. First, off-diagonal elements are in general small and will be neglected. We are therefore left with the $5f^{5/2}$ state split into two levels which are 2-fold (Γ_7), and 4-fold (Γ_8) degenerate, and with the $5f^{7/2}$ state split into three levels which are 2-fold (Γ_6), 2-fold (Γ_7) and 4-fold (Γ_8) degenerate. Second, because in plutonium the intermultiplet spin–orbit splitting is much larger than the intramultiplet crystal-field splitting ($>5:1$), we reduce the problem of solving AIM for the levels separately by treating the $5f^{5/2}$ Γ_7 and Γ_8 levels as one 6-fold degenerate level, and the $5f^{7/2}$ Γ_6 , Γ_7 and Γ_8 levels as another 8-fold degenerate level.

To solve the AIM in the general multi-orbital case, we use an interpolation scheme for its self-energy which is very accurate at both small and large frequencies, inspired by the success of the

iterative perturbation theory¹¹. At low frequencies, the exact value of the self-energy and its slope is extracted from the Friedel sum rule and from a slave-boson treatment¹⁶. At high frequencies, the self-energy is computed using self-consistent determination of moments¹⁷. The results of a smooth interpolation between these limits were in good agreement with quantum Monte Carlo data in the regimes where this comparison was possible (small degeneracy and high temperature).

The total energy of the system is evaluated by correcting the LDA total energy—which contains all the electrons including the core electrons—for the fact that it does not treat the f electrons correctly (in the same spirit as the LDA+U method; for a review, see ref. 18). We subtract from the LDA the average interaction energy of the f electrons (double counting term) and the average kinetic energy of the f electrons, and then add improved estimates of these quantities using the solution of the self-consistent AIM. The interaction term is evaluated using the Galitski–Migdal formula¹¹, and the improved f -electron kinetic energy within the DMFT method is extracted from the hybridization of the AIM¹¹. This gives a robust total energy scheme for practical applications.

To study the phase diagram of plutonium, we implement this self-consistent dynamical mean-field method. We include relativistic spin–orbit coupling effects, which are generally important for actinide compounds. The δ -phase of plutonium has a face-centred cubic (f.c.c.) structure with one atom per cell. We determine the ground-state energy as a function of the atomic volume V for the given lattice. Our theory needs a value of the effective Coulomb interaction U between f electrons in Pu. Various estimates exist in the literature. They indicate that the average interaction among f electrons is around 4 eV (ref. 19).

To illustrate the importance of correlations in our theory, we present results for values of U equal to 0 and 4 eV. First, the total energy as a function of volume is computed for $U = 0$ (GGA curve), which indicates a minimum at $V/V_\delta = 0.7$, (V_δ is the experimental volume of the δ -phase) close to the volume of the α -phase, Fig. 1. Certainly, we expect that correlations should be less important for the compressed lattice in general, but there is no sign of the δ -phase in the $U = 0$ calculation. The total energy curve is very different for $U > 0$. The details depend sensitively on the actual value of U . The curve at $U = 4$ eV shows the possibility of a double minimum; it is

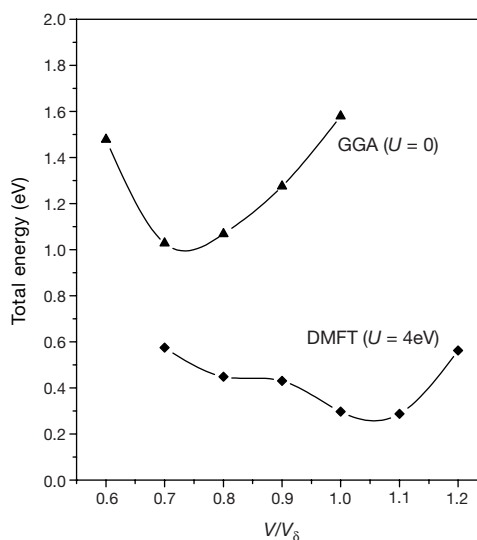


Figure 1 Calculated total energy for plutonium as a function of volume using dynamical mean-field theory (DMFT). The result of a density-functional GGA (generalized gradient approximation) calculation is also shown for comparison. Calculations were performed for the temperature 600 K, at which the δ -phase is stable. U denotes average Coulomb energy among f -electrons.

actually realized for a slightly smaller value of U . We can directly assign the small-volume feature to what would become the α -phase if we allowed for monoclinic distortions and a volume-dependent U , and assign the other feature to the δ -phase. As the energies are so similar, we may expect that as temperature decreases, the lattice will undergo a phase transition from the δ -phase to the α -phase with the remarkable decrease of volume by 25%. Our first-principles calculation contains the gist of the main features of the experimental phase diagram of plutonium.

The double-well behaviour in the total energy curve is unprecedented in LDA- or GGA-based calculations, but is a natural consequence of the proximity to a Mott transition. Indeed, recent studies¹¹ of model hamiltonian systems have shown that when the f -orbital occupancy is an integer and the electron–electron interaction is strong, two DMFT solutions, which differ in their spectral distributions, can coexist. It is natural that allowing the density to relax in these conditions can give rise to double minima, as seen in Fig. 1. This theoretical result deserves more detailed investigation, in the light of recent suggestions that the δ -phase might be metastable¹.

As a second result, we report our calculated spectral density of states for the δ -phase. We predict the appearance of a strong quasiparticle peak near the Fermi level. Recently, the δ -phase spectrum has been determined experimentally²⁰; our calculations are in accord with these measurements (Fig. 2). We note that LDA is not able to resolve a quasiparticle peak, and the electronic structure near the Fermi level falls into a deep minimum contrary to experiment. Even the LDA+ U theory^{21,22} fails to show the Kondo-like peak near the Fermi level. The DMFT description brings this feature into the spectrum, and the peak is quite narrow owing to both small hopping integrals and strong Coulomb interactions.

A simple physical explanation drawn from these calculations suggests that in the δ -phase the f -electron is slightly on the localized side of the interaction-driven localization–delocalization transition, with a sharp and narrow Kondo-like peak and well-defined upper and lower Hubbard bands. It therefore has the largest volume across all phases as has been found by previous GGA+ U

calculations^{21,22} that take into account Hubbard bands only. The low-temperature α -phase is more metallic; that is, it has larger spectral weight in the quasiparticle peak and smaller weight in the Hubbard bands. It will therefore have a much smaller volume that is eventually reproduced by LDA/GGA calculations which neglect both Coulomb renormalizations of quasiparticles and atomic multiplet structure. The delicate balance of the energies of the two minima may be the key to understanding the anomalous properties of Pu, such as the great sensitivity to small amounts of impurities (which would raise the energy of the less-symmetric monoclinic structure, thus stabilizing the δ -phase to lower temperatures) and the negative thermal expansion. We note, however, that the α -phase is not a weakly correlated phase: it is just slightly on the delocalized side of the localization–delocalization transition. This is a new view of plutonium; traditionally, the α -phase is regarded as well understood within LDA. However, the correlation viewpoint is consistent with a series of anomalous transport properties in the α -phase that are reminiscent of heavy-electron systems. For example, the resistivity of α -Pu around room temperature is anomalously large, temperature independent, and above the Mott limit²³ (the maximum resistivity allowed to the conventional metal). The same is true for the thermoelectric power²⁴. □

Received 21 December 2000; accepted 21 February 2001.

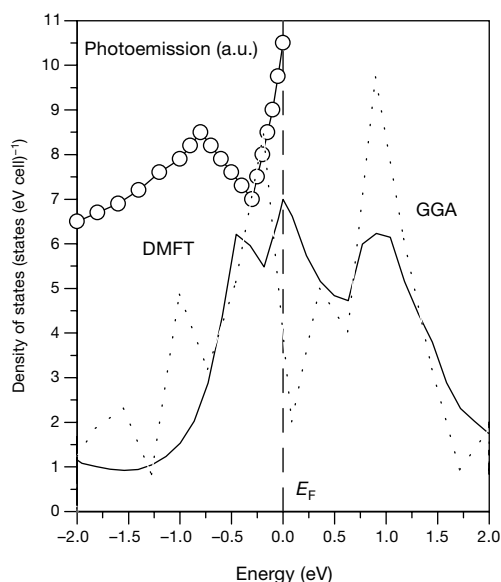


Figure 2 Comparison between calculated density of states for δ -plutonium using dynamical mean-field theory (DMFT; solid line) and recent photoemission experiments²⁰ (circles). The result of a density-functional GGA calculation is also shown for comparison (dotted line). Calculations are performed for the temperature 600 K at which the δ -phase is stable. E_F denotes the position of the Fermi level.

1. Broad, W. J. Russian plutonium research raises nuclear questions. *New York Times* 25 Jan. (2000).
2. Jeanloz, R. Science-based stockpile stewardship. *Phys. Today* **12**, 44–50 (2000).
3. Hecker, S. S. & Timofeeva, L. F. A tale of two diagrams. *Los Alamos Sci.* **26**, 244–251 (2000).
4. Freeman, A. J. & Darby, J. B. (eds) *The Actinides: Electronic Structure and Related Properties* Vols 1 and 2 (Academic, New York, 1974).
5. Johansson, B. The α - γ transition in cerium is a Mott transition. *Phil. Mag.* **30**, 469–479 (1974).
6. Lundqvist, S. & March, S. H. (eds) *Theory of the Inhomogeneous Electron Gas* (Plenum, New York, 1983).
7. Soderlind, P., Eriksson, O., Johansson, B. & Wills, J. M. Electronic properties of f -electron metals using the generalized gradient approximation. *Phys. Rev. B* **50**, 7291–7294 (1994).
8. Jones, M. D., Boettger, J. C., Albers, R. C. & Singh, D. J. Theoretical atomic volumes of the light actinides. *Phys. Rev. B* **61**, 4644–4650 (2000).
9. Solov'yev, I. V., Lichtenstein, A. I., Gubanov, V. A., Antropov, V. P. & Andersen, O. K. Spin-polarized relativistic linear-muffin-tin-orbital method: volume-dependent electronic structure and magnetic moment of plutonium. *Phys. Rev. B* **43**, 14414–14422 (1991).
10. Hubbard, J. Electron correlations in narrow energy bands. *Proc. R. Soc. Lond. A* **281**, 401–423 (1964).
11. Georges, A., Kotliar, G., Krauth, W. & Rozenberg, M. J. Dynamical mean-field theory of strongly correlated fermion systems and the limit of infinite dimensions. *Rev. Mod. Phys.* **68**, 13–125 (1996).
12. Anisimov, V. I., Poteryaev, A. I., Korotin, M. A., Anokhin, A. O. & Kotliar, G. First-principles calculations of the electronic structure and spectra of strongly correlated systems: dynamical mean-field theory. *J. Phys. Condens. Matter* **35**, 7359–7367 (1997).
13. Nekrasov, I. A. et al. Calculation of photoemission spectra of the doped Mott insulator $\text{La}_{1-x}\text{Sr}_x\text{TiO}_3$ using LDA+DMFT(QMC). *Eur. Phys. J. B* **18**, 55–61 (2000).
14. Anderson, P. W. Localized magnetic states in metals. *Phys. Rev.* **124**, 41–53 (1961).
15. Andersen, O. K. Linear methods in band theory. *Phys. Rev. B* **12**, 3060–3083 (1975).
16. Fressard, R. & Kotliar, G. Interplay of Mott transition and ferromagnetism in the orbitally degenerate Hubbard model. *Phys. Rev. B* **56**, 12909–12915 (1997).
17. Nolting, W. & Borgiel, W. Band magnetism in the Hubbard model. *Phys. Rev. B* **39**, 6962–6978 (1989).
18. Anisimov, V. I., Aryasetiawan, F. & Lichtenstein, A. I. First-principles calculations of the electronic structure and spectra of strongly correlated systems: the LDA+ U method. *J. Phys. Condens. Matter* **9**, 767–808 (1997).
19. Desclaux, J. P. & Freeman, A. J. in *Handbook on the Physics and Chemistry of the Actinides* (eds Freeman, A. J. & Lander, G. H.) Vol. 1 (Elsevier, Amsterdam, 1984).
20. Arko, A. J., Joyce, J. J., Morales, L., Wills, J. & Jashley, J. Electronic structure of α - and δ -Pu from photoelectron spectroscopy. *Phys. Rev. B* **62**, 1773–1779 (2000).
21. Savrasov, S. Y. & Kotliar, G. Ground state theory of δ Pu. *Phys. Rev. Lett.* **84**, 3670–3673 (2000).
22. Bouchet, J., Siberchicot, B., Jollet, F. & Pasturel, A. Equilibrium properties of δ -Pu: LDA+ U calculations (LDA equivalent to local density approximation). *J. Phys. Condens. Matter* **12**, 1723–1733 (2000).
23. Boring, A. M. & Smith, J. L. Plutonium condensed-matter physics—a survey of theory and experiment. *Los Alamos Sci.* **26**, 91–127 (2000).
24. Lallement, R. Dilatation et pouvoir thermoélectrique du plutonium α a basse température. *J. Phys. Chem. Solids* **24**, 1617–1624 (1963).

Acknowledgements

We thank A. Lichtenstein for discussions. This work was supported by the DOE division of Basic Energy Sciences and by Los Alamos National Laboratory.

Correspondence and requests for materials should be addressed to S.Y.S. (e-mail: savrasov@physics.rutgers.edu).

Optical polymer thin films with isotropic and anisotropic nano-corrugated surface topologies

Mohammed Ibn-Elhaj & Martin Schadt

ROLIC Research Ltd, Gewerbestrasse 18, 4123 Allschwil, Switzerland

Light reflection from computer monitors, car dashboards and any other optical surface can impair the legibility of displays, degrade transmission of optical components and in some cases may even pose safety hazards. Antireflective coatings are therefore widely used, but existing antireflection technologies often perform sub-optimally or are expensive to implement. Here we present an alternative approach to antireflection coatings, based on an extension of our photo-aligning and photo-patterning technology for liquid-crystal displays^{1,2} (LCDs) and liquid-crystal polymer films with smooth surfaces^{3,4} to optical polymer films with controlled surface topologies. Nano- and micro-corrugated topologies are shown to result from optically induced monomer phase-separation on the polymer surfaces. The properties of the resulting films make them suitable high-performance and low-cost antireflection coatings for optical components of virtually any size, shape and material. Moreover, the approach can be used to form a wide range of other functional polymer thin films with isotropic as well as anisotropic topologies. For example, films can be produced whose optical birefringence exceeds that of the birefringence of the polymer material itself. These new films can also be used as diffractive thin films, diffusers, and directional reflectors which preserve light polarization, or as substrates for aligning liquid crystals to produce bright, low-power-consumption LCDs with integrated optical functions and memory.

Antireflective coatings are needed to realize highly transmissive optical components made of glass, plastic, silicon, and so on, or to achieve high contrast and brightness in displays or instrument panels. There exists no single antireflection technology meeting the requirements of all applications with their diverse sizes, shapes and materials. Apart from vacuum-deposited dielectric thin films⁵, virtually all antireflection techniques are based on the production of densely packed nanopores on the surface(s) of optical components. If the dimensions of the nanopores are made smaller than the wavelength of light, no light scattering occurs and porous films act in an antireflective fashion. The nanopores reduce the effective refractive index n_f of the film such that it meets the antireflection criterion⁵ $n_f = (n_a n_s)^{0.5}$, where the respective indices n_a and n_s are those of air and substrate at a given wavelength.

Examples of isotropic surfaces with isotropic nanopores that are antireflective are etched surfaces⁶, sol-gel derived coatings⁷, randomly deposited particles⁸, and arrays of pyramids formed by photochemical or mechanical means^{9,10}. Anisotropic surface topologies in films made of optically isotropic materials—such as linear gratings or grooves—exhibit directional antireflectivity¹¹, and have been shown to align liquid crystals^{12,13} or to act as phase-retarders¹⁴. However, existing high-resolution topologies are expensive or difficult to apply to large or curved areas.

De-mixing of binary blends of mutually incompatible polymers during coating, and subsequent removal of one of the polymer components, generates nanopores in the residual film; the shape and distribution of these nanopores are isotropic, and the residual film exhibits excellent antireflective properties¹⁵. However, because the dimensions of the pores are determined by the de-mixing time of the two polymer components, pore size is difficult to control. Moreover, owing to the inherently inert character of processed polymers,

the adhesion of the film and its interaction with substrates is weak. This may lead to polymer contraction within the film and to the formation of polymer islands (de-wetting¹⁶), accompanied by deteriorating antireflectivity. Because of the random distribution of the nanopores, neither geometrical nor optical anisotropy can be achieved with films made by polymer de-mixing.

We have shown^{3,4,17–21} that a broad range of optically anisotropic thin-film devices with smooth surfaces may be produced by aligning, photo-patterning and/or stacking surface-aligned liquid-crystal polymer films on single substrates of virtually any material. We have started to investigate whether corrugations^{22,23} can be generated at the surface of polymer films which exhibit optically controllable and stable topologies. We report here that phase-separation can be optically induced in monomer liquid-crystal films deposited on single substrates such that nano- and micro-corrugated surfaces result. Moreover, we show that substrate alignment determines the shape and distribution of the corrugations, enabling the production of unusual optical, mechanical and electrical functions. Unlike the technique of Walheim *et al.*¹⁵, our monomer-corrugation (MC) technology is based on blends of fully miscible and thermodynamically stable monomers—such as blends of liquid-crystal pre-polymers and monomeric liquid crystals (LCs)—where only the pre-polymers are photo-crosslinkable. After coating with an MC film, phase-separation is induced by crosslinking the pre-polymer-component(s) with exposure to ultraviolet radiation. Depending on the design of the MC mixture and details of film preparation, subsequent removal of LC component(s) leaves nano- or micro-pores in the crosslinked, solid polymer film. Figure 1a shows an atomic force microscope (AFM) image and the profile of the nanopore surface topology of a film produced in this manner on a plastic substrate. The substrate does not exhibit surface alignment. The shape and the distribution of the pores are isotropic.

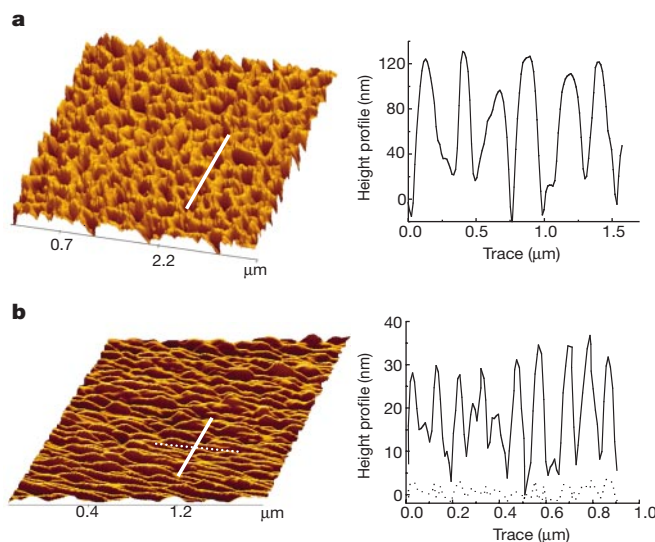


Figure 1 Geometrically isotropic and anisotropic MC nano-topologies. **a**, AFM image and profile of the isotropic nano-topology of a monomer-corrugated (MC) film on a non-aligned glass substrate. The average diameter and height of the pores are 200 nm and 120 nm, respectively. Solid white line, location of AFM profile. **b**, AFM image and profile of the anisotropic nano-topology of an MC liquid crystal polymer (LCP) film induced by an underlying, uniaxial alignment layer on glass. The aligning direction of the 50-nm, LPP photo-alignment layer parallels the dotted white line. Solid white line, location of AFM profile. The MC mixture used in all experiments consists of the photo-crosslinkable LCP mixture CB483 and the LC component R03952 (of ROLIC). The LPP photo-aligning mixture is JP265 (of ROLIC). All materials have been developed by ROLIC, and some are commercialized via Vantico. Refractive indices of CB483 at 500 nm: $\Delta n_{LCP} = (n_{b,LCP} - n_{o,LCP}) = 0.12$; $n_{o,LCP} = 1.566$ and $n_{b,LCP} = 1.686$. The average period and height of the MC nano-grooves are $\Lambda = 80$ nm and $h = 30$ nm, respectively.

Liquid crystal pre-polymers and LCs can be uniaxially aligned on single substrates²¹. In the case of optically induced alignment, the alignment can in addition be photo-patterned by means of photo-masks^{3,4,20}. To achieve not only isotropic pores (as in Fig. 1a) but also anisotropic MC topologies requires laterally anisotropic aligning forces in the MC films which are strong enough to induce anisotropic phase-separation. To generate this type of anisotropy, we decided to investigate mechanically aligned or photo-aligned substrate surfaces^{3,4,21}. To avoid defects and contamination, and to enable LC alignment patterns, we used our contact-free linear photo-polymerization (LPP) technology^{1,2} for substrate alignment. After processing and exposure of an LPP layer to linearly polarized ultraviolet radiation, LC molecules in contact with this layer align, with their long axes parallel to the direction of polarization of the radiation^{1,2}. Figure 1b shows an AFM image and the profile of a nematic liquid-crystal polymer (LCP) thin-film topology which results when we apply our MC process to a glass plate with uniaxial surface alignment. From Fig. 1b it follows that the shape of the MC nanopores and their direction are indeed strongly anisotropic, with the elongated pores aligned parallel to the underlying alignment direction of the substrate. Analogous anisotropic MC thin-film topologies can be achieved on non-aligned substrates by using MC phase-separation mixtures which do not require underlying alignment²⁴.

The AFM images in Fig. 1 show that monomer corrugation by phase-separation in polymer films on single substrates not only enables uniform isotropic nano-topologies but also—to our knowledge, for the first time—anisotropic nano-topologies. Our MC technology is applicable to almost any planar or non-planar substrate or material. The amplitude and periodicity of the nano-

topology are tunable by adjusting the film thickness, the LC-pre-polymer content, the aligning direction of the substrate, and the energy of the crosslinking radiation. This allows us to control the period and height of the vertical groove profile, as well as the distribution and volume ratio of the grooves. We show below that nano- and micro-corrugated MC films can be used to produce a broad range of optical thin-film elements. For example, broadband antireflective coatings, optically strongly birefringent LC polymer films, conducting and bright diffuse reflectors, optical gratings and topological aligning layers for LCDs may be produced. By combining the elastic LC aligning forces of periodic surface topologies²⁵ with the van der Waals forces of photo-alignment^{1,26}, hybrid alignment results.

The transmission of an ordinary, flat sheet of glass in air is typically less than 92%; that is, more than 4% of incident white light is reflected at each of the two glass/air interfaces. This seriously degrades the transmission of optical components, and also degrades the visibility of pictures, text or other optical information behind glass, plastic panels or LCD polarizers. Isotropic MC nano-topologies like those shown in Fig. 1a act as efficient antireflection coatings. This is illustrated in Fig. 2a, which is a photograph of white text on a black background behind a sheet of glass consisting of differently reflecting segments. Segment (3) illustrates the excellent antireflection performance of our MC films which render glass virtually invisible when applied to both glass surfaces. The legibility of the text behind glass segment (2) is still considerably better than that behind the untreated glass segment (1) whose reflectivity almost buries the text. The visual impression of Fig. 2a is confirmed

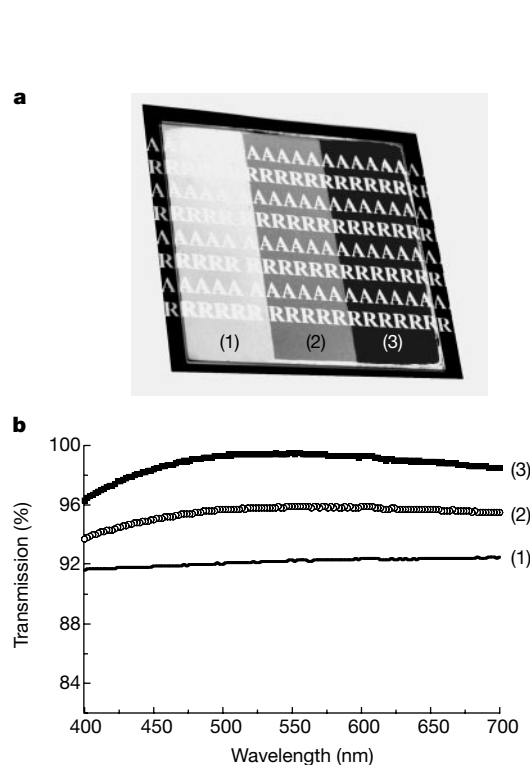


Figure 2 Highly efficient, isotropic, broadband antireflective MC coatings on glass. **a**, Photograph of a white text on a black background behind a sheet of glass with different reflectivities of the glass in the three segments. (3) Both glass surfaces coated with a geometrically isotropic nano-corrugated MC film; (2) only front surface coated with MC film; (1) bare glass with no antireflection coatings. **b**, Dependence of the transmission of the three glass segments of **a** versus wavelength of vertically incident light. Segment (3) exhibits a very low reflectivity of only 0.1% in the green and a high integral transmission $T > 99.1\%$ over the entire visible spectral range from 400 to 700 nm.

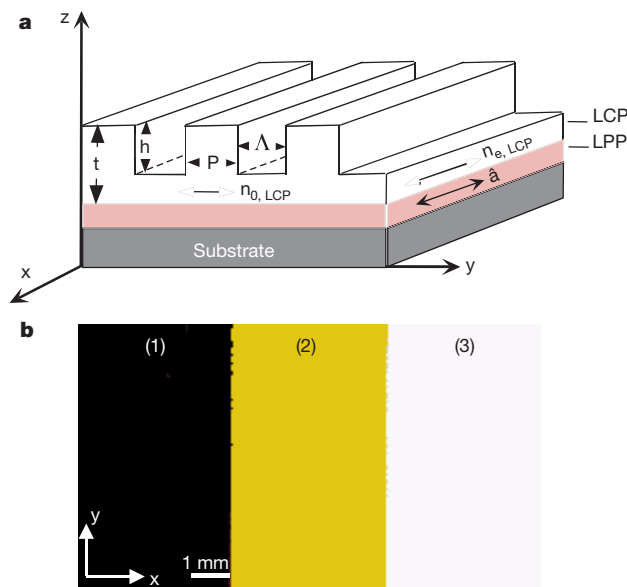


Figure 3 Anisotropic MC topologies on aligned substrates and hybrid MC alignment. **a**, Schematic of topologically and optically anisotropic MC LCP film; $\hat{a}$, substrate aligning direction; $f = \Lambda/(P + \Lambda)$, volume fraction of MC-groove topology. **b**, Photograph of a transmissive nematic LCP film on a glass substrate (white back-light) between crossed polarizers. Section (1): the LCP-film is MC-aligned and the LCP film thickness is chosen such that the film acts as a half-wave retarder for white light between crossed polarizers (that is, optical retardation $\delta(LCP) = 230$ nm). Exit polarizer, parallel to y axis; MC groove geometry, $\Lambda \approx 200$ nm, $h \approx 30$ nm. Sections (2) and (3): the same LCP film as in (1) but aligned by the combined aligning forces of the MC topology and a superimposed LPP photo-aligning layer (not shown in **a**). The LPP film is photo-aligned at 45° with respect to the direction of the MC grooves before spin-coating the LCP retarder onto the substrate. The relative azimuthal aligning anchoring strengths in the two sections are: $LPP(2) \approx MC(1)$ and $LPP(3) \gg MC(1)$. The different anchoring strengths of the LPP layers are due to different LPP UV-exposure times²⁷ (6 and 12 s at 2 mW cm^{-2}).

by the wavelength dependence of the transmission of the three glass segments of which segment 3 exceeds 99.1% over the entire visible spectral range (Fig. 2b). This broadband antireflection performance of our MC coatings is due to the continuously narrowing pore diameters (from top to bottom in Fig. 1) and to their low effective refractive index of refraction $n_{\text{eff}} = 1.1$, which approaches the index of air.

The anisotropic nano-topologies generated by our MC process can also act as aligning layers for LC molecules. Figure 3a schematically depicts an MC nano-topology whose grooves copy the underlying uniaxial alignment direction $\hat{a}$ of the substrate. As the size and the depth of the MC grooves can be varied, deep grooves with small radii of curvature and correspondingly large LC anchoring energies can be achieved. To demonstrate LC alignment by MC nano-topologies of the type depicted in Fig. 1b, we chose the demanding case of alignment of a nematic LCP film on a single substrate comprising an anisotropic MC coating. As in our earlier work^{3,4,21}, the LC-pre-polymer film was spin-coated onto the MC layer and then crosslinked. Segment (1) in Fig. 3b shows an MC aligned LCP half-wave retarder between crossed polarizers. The uniform black appearance confirms that the LCP film indeed uniaxially aligns parallel to the nano-grooves of the MC-coated substrate, and that alignment does not deteriorate during cross-linking.

To investigate whether hybrid alignment between MC- and photo-alignment is possible, we coated the MC aligning layer in sections (2) and (3) of Fig. 3b with an LPP photo-aligning layer^{1,2} whose azimuthal anchoring energy²⁷ in segment (2) is smaller than in (3), and whose aligning direction deviates by 45°. From the fully transmissive white section (3) in Fig. 3b it follows that the strong anchoring of photo-alignment in this section completely overrides MC alignment and causes the slow LCP retarder axis to rotate by 45°. In section (2) where the anchoring strengths of MC- and photo-alignment are comparable, the half-wave retarder appears yellow. This is due to the rotation of the slow optical LCP-retarder axis by 22°—that is, between the directions of MC- and photo-alignment—and proves that photo- and MC-alignment can indeed be combined to give hybrid MC/photo-alignment. A possible application would be in hybrid aligned optically bistable LCDs. Such displays memorize their last switching state after disconnecting their power supply²⁸.

Unlike the case of LC polymer films with smooth surfaces, we

now show that the effective optical retardation δ_{eff} of (birefringent) LCP films with anisotropic nano-topologies depends not only on the film thickness and birefringence Δn_{LCP} of the LCP material, but also on the degree of anisotropy of the pore topology. When using the simple topology of Fig. 3a as a model for MC LCP retarder films, and extending the formalism for optically isotropic materials in ref. 14 to our anisotropic case, we obtain a rough estimate for the δ_{eff} value of LCP films with anisotropic nano-topologies:

$$\delta_{\text{eff}} = \delta_{\text{LCP}} + \delta_{\text{topology}}$$

$$= \Delta n_{\text{LCP}}(t - h) + h\{[f + n_e^2(1 - f)]^{0.5} - n_0[n_0^2f + (1 - f)]^{-0.5}\}$$

For a conventional, uniaxially aligned LCP film with smooth surface ($\Lambda = 0$), equation (1) yields $\delta_{\text{eff}} = \delta_{\text{LCP}} = \Delta n_{\text{LCP}}t$; whereas in the case of a periodically corrugated isotropic dielectric ($n_e = n_0$) the birefringent first term in equation (1) vanishes, leaving only the topological term δ_{topology} (ref. 14). From equation (1) it follows that LCP films made of weakly birefringent materials which exhibit anisotropic nano-topologies are still strongly optically retarding. Our MC technology indeed enables the production of strongly birefringent LCP films based on low-birefringent LCP materials (Fig. 4). Owing to its anisotropic nano-topology, the increase of the effective birefringence of the MC film in Fig. 4 exceeds δ_{LCP} by 20%. Because strongly birefringent LCP materials tend to be optically unstable, this finding is important for applications in bright light environments (such as projectors).

The nano-topologies of the above MC thin-film devices are all designed such that no light scattering occurs in the visible range. To achieve scattering, the pore topology must be scaled into the micrometre range. This is demonstrated in Fig. 5b, which shows two AFM images of a MC LCP film on glass with two different photo-aligning directions^{1–3}. Anisotropic MC micro-topologies of the type depicted in Fig. 5b act as efficient directional light-diffusers. This is shown in Fig. 5a by the photograph of a photo-aligned and photo-patterned micro-MC film which is phase-separated on an aluminium mirror and therefore acts as a diffuse, directional reflector. The two LPP alignment directions of the digital picture elements in Fig. 5a and the two resulting slow optical axes of the

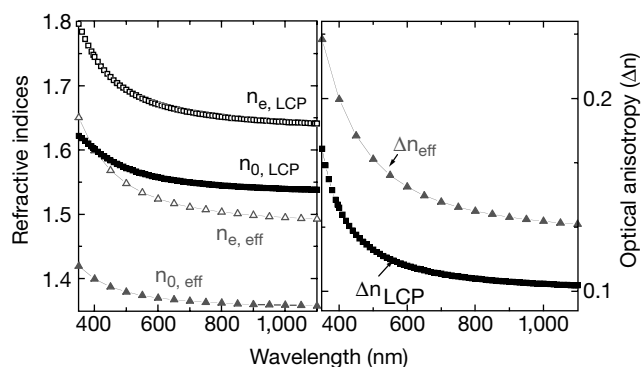


Figure 4 Optically strongly birefringent MC film made with low-birefringence LCP. The figure shows effective refractive indices and optical anisotropy versus wavelength of incident light of an MC LCP retarder film made of the low birefringence LCP material CB483. MC film geometry: $\Lambda \approx 150$ nm, $h \approx 75$ nm and $t \approx 110$ nm (compare Fig. 3a). Left, the respective wavelength dependence of the effective refractive indices of the LCP film with anisotropic nano-topology ($n_{0,\text{eff}}$ and $n_{e,\text{eff}}$) as well as the intrinsic refractive indices of the LCP material ($n_{0,\text{LCP}}$ and $n_{e,\text{LCP}}$). Right, the corresponding effective optical anisotropy Δn_{eff} of the MC LCP film and the intrinsic anisotropy Δn_{LCP} of the LCP material.

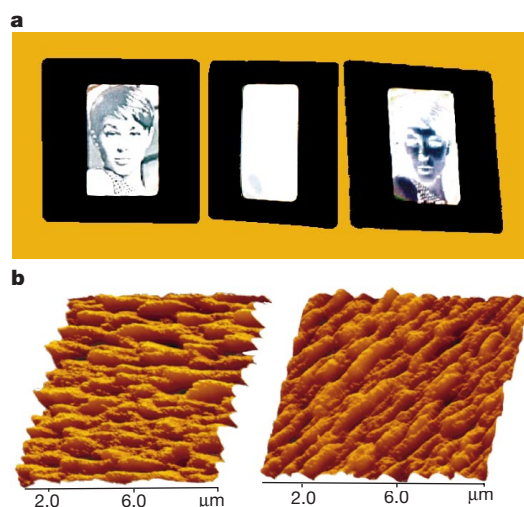


Figure 5 Bright directional MC thin-film diffusers/reflectors. **a**, Viewing-angle dependence of a photo-aligned and photo-patterned micro-corrugated LCP thin film coated on an aluminium mirror. The digital picture elements of the diffuse and directionally reflecting MC film consist of the two micro-topology directions shown in **b**. MC film geometry: $\Lambda \approx 900$ nm, $h \approx 80$ nm and $t \approx 400$ nm. **b**, AFM images of the micro-topologies of two pixels of the MC film which are uniaxially photo-aligned under 45°. MC film thickness, 400 nm.

pixelated MC topology differ by 45° (compare Fig. 5b and ref. 29). Index mismatch along a particular MC direction for incident light which is polarized along this direction leads to scattering and reduced transmission, whereas no scattering occurs along directions with no index mismatch. Owing to the large refractive index differences between the pores (air) and the corrugated LC polymer parts of our MC films, large index-mismatching can be achieved (~ 0.5) for specific viewing angles. This is illustrated by the central image in Fig. 5a, which disappears because it is taken at an angle where the intensities of the two polarization states of the reflected light are equal. Conversely, the left and the right photographs in Fig. 5a are taken at viewing angles where the pixelated MC topology scatters light in such a way that either the positive or the negative scattering image appears. The topologically and optically anisotropic MC micro-topologies which generate the bright images in Fig. 5 show that our MC technology enables the production of highly efficient directional thin-film diffusers and reflectors. By combining MC topologies with conducting properties—for instance, via metallic top-coatings or conducting MC polymers—their optical functions can be combined and integrated into displays to increase brightness and functionality.

Optically isotropic or anisotropic MC films may be applied to almost any substrate, independent of size or shape. They enable the production of a broad range of novel optical thin-film devices, such as broadband antireflection and antiglare coatings. Moreover, the surface topologies of MC films made of LC polymers can be photo-patterned and combined with LCP films with smooth surfaces^{18,21}. This opens up such diverse applications as aligning layers for LCDs, diffuse and directional reflectors, optically strongly anisotropic retarders consisting of low birefringent materials, optical security elements, reflective and transmissive polarizers, diffractive optical thin films, as well as polarization converters. Our MC technology enables the integration of optical, electrical and aligning functions in hybrid thin films, and has the potential to produce bright, flat panel displays. □

Received 6 November 2000; accepted 23 January 2001.

- Schadt, M., Schmitt, K., Kozinkov, V. & Chigrinov, V. Surface-induced alignment of liquid crystals by linearly polymerized photopolymers. *Jpn J. Appl. Phys.* **31**, 2155–2164 (1992).
- Schadt, M., Seiberle, H. & Schuster, A. Optical patterning of multi-domain liquid-crystal displays with wide viewing-angles. *Nature* **381**, 212–215 (1996).
- Schadt, M., Seiberle, H., Schuster, A. & Kelly, S. M. Photo-induced alignment and patterning of liquid crystalline polymer films on single substrates. *Jpn J. Appl. Phys.* **34**, 3240–3249 (1995).
- Benecke, C., Seiberle, H. & Schadt, M. Determination of director distributions in liquid crystal polymer-films by means of generalized anisotropic ellipsometry. *Jpn J. Appl. Phys.* **39**, 525–531 (2000).
- Bass, M. (ed.) *Handbook of Optics. Fundamentals, Techniques & Design* (McGraw-Hill, New York, 1995).
- Fraunhofer, J. *Versuche ueber die Ursachen des Anlaufens und Mattwerdens des Glases und die Mittel, denselben zu vorzuzukommen* (Gesammelte Schriften, München, 1888).
- Thomas, I. M. Method for preparing porous silica antireflection coatings varying in refractive index from 1.22 to 1.44. *Appl. Opt.* **31**, 6145–6149 (1992).
- Moulton, H. M. Method of forming a reflection reducing coating. US Patent No. 2536764 (1951).
- Clapham, P. B. & Hutley, M. C. Reduction of lens-reflection by the 'moth eye' principle. *Nature* **244**, 281–282 (1973).
- Heine, C. & Morf, R. H. Submicron gratings for solar applications. *Appl. Opt.* **34**, 2476–2482 (1995).
- Hodgkinson, I. J. Linear and circular birefringence of coatings fabricated by serial bi-deposition. *Proc. SPIE* **3790**, 119–132 (1999).
- Cognard, J. Alignment of nematic liquid crystals and their mixtures. *Mol. Cryst. Liq. Cryst.* **78** (Suppl. 1), (1981).
- Rütschi, M., Grüter, P., Fünfschilling, J. & Güntherodt, H.-J. New method for creating liquid crystal devices. *Science* **265**, 512–514 (1994).
- Flanders, D. C. Submicrometer periodicity gratings as artificial anisotropic dielectrics. *Appl. Phys. Lett.* **42**, 492–494 (1983).
- Wallheim, S., Schäffer, E., Mlynjek, J. & Steiner, U. Nanophase-separated polymer films as high-performance antireflection coatings. *Science* **283**, 520–522 (1999).
- Reiter, G. Unstable thin polymer films: rupture and de-wetting processes. *Langmuir* **9**, 1344–1351 (1993).
- Chen, K., Chang, C., DelPico, J., Seiberle, H. & Schadt, M. Integral-color plastic LCD: the ultimate flat-panel display. *Proc. Soc. Inform. Displ. SID99* 1054–1057 (1999).
- Schadt, M. Liquid crystal displays and novel optical thin-films enabled by photo-alignment. *Proc. Int. Liq. Cryst. Conf. ILCC2000* (in the press).
- Schmitt, K., Fünfschilling, J., Cherkaoui, Z. & Schadt, M. Fast time-sequential color display based on cholesteric filters and DHF-LCDs. *Proc. Int. Displ. Res. Conf. IDRC99* 437–440 (1999).

- Moia, F., Seiberle, H. & Schadt, M. Optical LPP/LCP-devices: novel optical security elements. *Proc. SPIE* **3973**, 196–303 (2000).
- Schadt, M., Schuster, A. & Seiberle, H. Optical component and method of manufacture. US Patent No. 6160597 (1994).
- Virga, E. G. & Schadt, M. Corrugations on the free surface of liquid crystal layers. *Jpn J. Appl. Phys.* **39**, 6637–6642 (2000).
- Virga, E. G. Disclinations and bumps in liquid crystals. *Phil. Trans. R. Soc. Lond. A* **355**, 2035–2044 (1997).
- Seiberle, H. & Schadt, M. Liquid crystal photo-alignment layer. Int. Patent Application PCT/IB99/01001 (1998).
- Bereman, D. W. Surface shape and alignment of nematic liquid crystals. *Phys. Rev. Lett.* **28**, 1683–1686 (1972).
- Gibbons, W. M., Shannon, P. J., Sun, S. T. & Swetlin, B. J. Surface-mediated alignment of nematic liquid crystals with polarized laser light. *Nature* **351**, 49–50 (1991).
- Schadt, M. & Seiberle, H. Optical patterning of multidomain LCDs. *Proc. Soc. Inform. Displ. SID97* 397–400 (1997).
- Jones, J. C. *et al.* Zenithal bistable devices: towards the electronic book with a simple LCD. *Proc. Soc. Inform. Displ. (Int. Displ. Workshop 2000) SID IDW00* 301–304 (2000).
- Schadt, M. Advances in LPP-photo-alignment of liquid crystals applied to the phase-retarder image of Alfred Saupe. *Mol. Cryst. Liq. Cryst.* **292**, 235–243 (1997).

Correspondence and requests for materials should be addressed to M.S.
(e-mail: martin.schadt@rolic.ch).

Removal of chlorofluorocarbons by increased mass exchange between the stratosphere and troposphere in a changing climate

Neal Butchart & Adam A. Scaife

The Met Office, London Road, Bracknell, Berkshire RG12 2SZ, UK

Chlorofluorocarbons (CFCs), along with bromine compounds, have been unequivocally identified as being responsible for most of the anthropogenic destruction of stratospheric ozone¹. With curbs on emissions of these substances, the recovery of the ozone layer will depend on their removal from the atmosphere. As CFCs have no significant tropospheric removal process, but are rapidly photolysed above the lower stratosphere, the timescale for their removal is set mainly by the rate at which air is transported from the troposphere into the stratosphere². Using a global climate model we predict that, in response to the projected changes in greenhouse-gas concentrations during the first half of the twenty-first century, this rate of mass exchange will increase by 3% per decade. This increase is due to more vigorous extra-tropical planetary waves emanating from the troposphere. We estimate that this increase in mass exchange will accelerate the removal of CFCs to an extent that recovery to levels currently predicted for 2050 and 2080 will occur 5 and 10 years earlier, respectively.

The effects of increasing greenhouse-gas concentrations on the transport between the troposphere and stratosphere have not previously been quantified in any stratospheric climate change experiment. Here we analyse the trend in this transport in 60-year runs³ of a troposphere–stratosphere version⁴ of The Met Office Unified Model⁵. This model has a latitude–longitude resolution of 2.5° by 3.75° , 49 levels extending from the surface to the 0.1-hPa level in the atmosphere (~ 64 km altitude) and a comprehensive set of physical parametrizations for sub-grid-scale processes. Orographic gravity-wave drag⁶ was included up to the 20-hPa level, and Rayleigh friction was applied in the mesosphere. Heating rates included the effects of minor greenhouse gases— CH_4 , N_2O , CFC11 (CCl_3F) and CFC12 (CCl_2F_2), as well as CO_2 (ref. 3).

Starting from conditions representative of the early 1990s, the model was run for the 60 years 1992 to 2051 (run A) assuming the Intergovernmental Panel on Climate Change IS92a scenario for greenhouse-gas concentrations, but no changes in the concentrations of stratospheric ozone or water vapour³. Sea surface conditions were taken from a separate coupled ocean–atmosphere experiment using the same greenhouse-gas scenario⁷. The run was repeated (run B) using different initial conditions, but also from the early 1990s, and the two runs were used to show the presence of unpredictable decadal variability. Nonetheless, both runs predicted a systematic global cooling of the stratosphere, which increased with height, and was broadly consistent with past observed trends³. Similarly, the tropospheric trends were consistent with those predicted by other climate models⁸.

Between the troposphere and stratosphere, a net upward mass flux across the tropical tropopause is balanced by a net downward mass flux in the extra-tropics⁷. Over the 60 years of both runs, there was no systematic change in the position of the ‘turnaround latitudes’ where, in the lower stratosphere, the seasonal- and annual-mean mass fluxes change direction from upwards to downwards (see Fig. 1a for run B). In the 1990s, the modelled upward mass fluxes (Fig. 1b) were in broad agreement with mass fluxes derived from observations⁹. At the same time, the mean vertical velocity from 12°N to 12°S at 68 hPa varied from ~0.16 mm s⁻¹ in June–August to ~0.3 mm s⁻¹ in December–February, slightly less than the 0.2–0.4 mm s⁻¹ inferred from the so-called ‘tape recorder’ signal in water vapour measurements^{10,11}: this signal results from the annual cycle in the tropical tropopause temperatures caused, in

turn, by the annual cycle in the upwelling¹². Again in the 1990s, the modelled tropical tropopause temperatures had the correct amplitude annual cycle and, notably, there was no change in the amplitude over the 60 years of the run, despite the predicted cooling of the lower stratosphere. Consistently, every season shows similar, large positive trends in tropical upwelling (Fig. 1b), with the net effect in run A of a 19% increase in the annual-mean mass flux over the 60 years, and a similar 18% increase in run B (Fig. 1b). In general, the concomitant trends in the residual mean meridional circulations^{13,14} agree qualitatively with those reported for doubled CO₂ experiments^{15,16}.

The main reason for the trends in our model was increased extra-tropical planetary-wave driving. By ‘downward control’¹⁷, the wave-induced mean zonal forces above the lower stratosphere in the model imply a downward mass flux in the extra-tropics which agrees quantitatively with the flux required to balance the tropical upwelling (compare Fig. 1a and d). The downward control mass fluxes are slightly larger, partly because of the absence of a small negative contribution from the Rayleigh friction due to easterly zonal mean winds in the mesosphere at the turnaround latitudes (Fig. 1c). Also, the missing contribution from orographic gravity-wave drag (which was not archived from our runs) probably causes the turnaround latitudes to be closer to the equator than those for the actual mass fluxes, especially in the Northern Hemisphere¹⁸ (compare Fig. 1a and c). Nonetheless, the 60-year increase of $6.4 \times 10^8 \text{ kg s}^{-1}$ in the wave-driven, annual-mean, downward mass flux balances most of the $8.7 \times 10^8 \text{ kg s}^{-1}$ increase in the tropical upwelling, confirming that the increase in mass transport

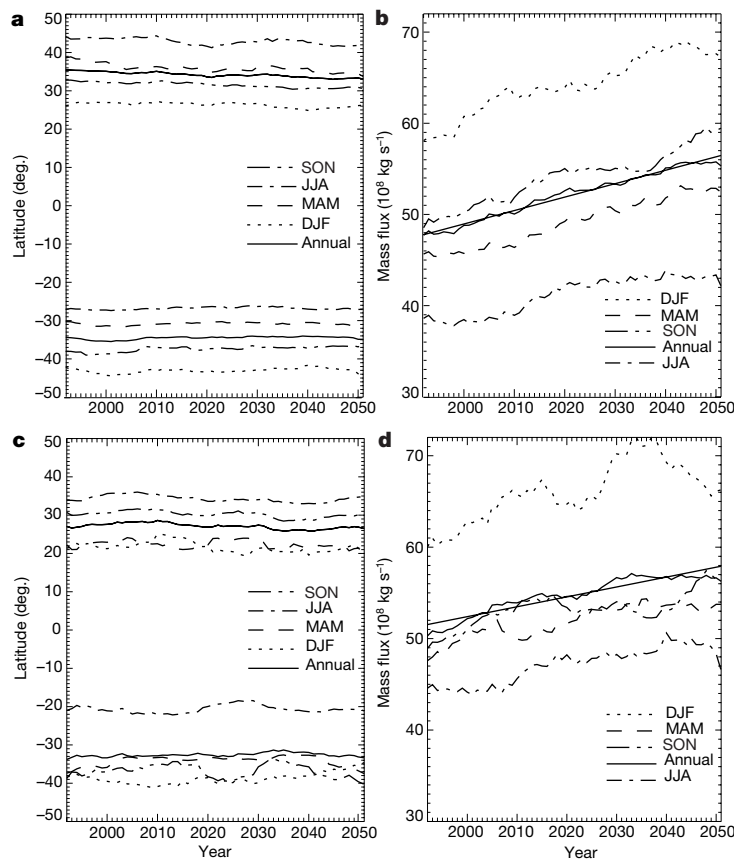


Figure 1 Mass exchange between the troposphere and stratosphere. **a**, Turnaround latitudes where the modelled residual vertical velocities^{13,14}, and hence vertical mass fluxes, change sign, and **b**, the integrated upward mass flux between the turnaround latitudes. **c**, Turnaround latitudes for the vertical mass fluxes derived from the planetary wave driving by downward control¹⁷ and, **d**, the integrated downward mass flux between

the poles and the turnaround latitudes in **c** obtained from the planetary wave driving assuming downward control. Runs A and B gave similar results, and curves show 11-year moving means at 68 hPa (~18.7 km) for run B only. The straight lines in **b** and **d** are least-squares fits to the annual-mean results.

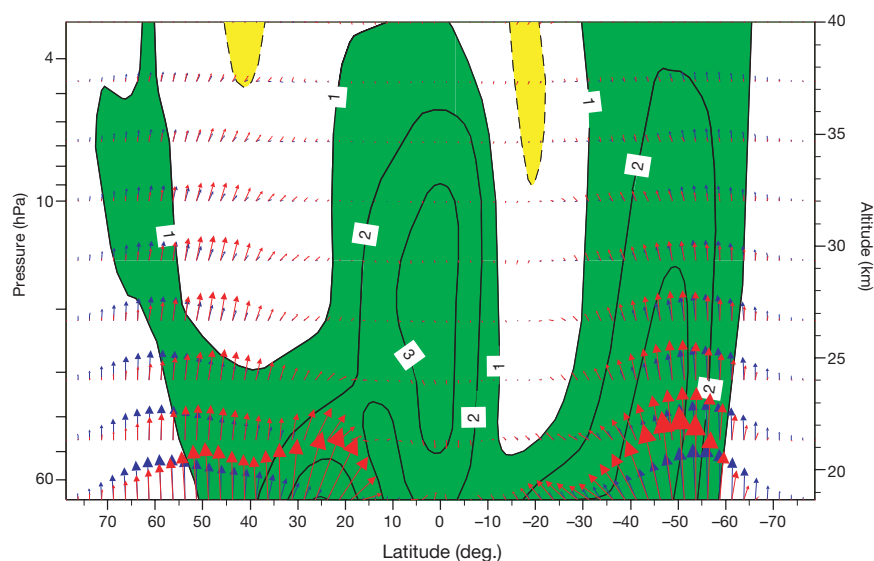


Figure 2 Predicted changes in planetary-wave flux. The figure shows the annual-mean Eliassen-Palm flux²⁸ of planetary-wave activity at the start of run B (blue arrows) and the change ($\times 10$) over the 60 years of the simulations (red arrows). The flux vectors have been multiplied by the cosine of latitude to give the natural form for plotting in the latitude–height plane²⁹; they have been further scaled such that the distance occupied by 10° of

latitude represents a value of $1.33 \times 10^7 \text{ kg s}^{-2}$, and that occupied by 10 km in altitude represents a value of $1.2 \times 10^5 \text{ kg s}^{-2}$. Also shown is the change in the zonal-mean zonal wind in m s^{-1} . Green shading denotes a westerly shift, and yellow, an easterly shift. The zero contour has been omitted. Results shown here were taken from a least-squares linear fit to the annual-mean data.

between the troposphere and stratosphere was predominantly wave-driven in the model.

Climate change produced, in run B, a westerly shift in the zonal-mean zonal winds in the tropics and, where the winds are rather weak, in the sub-tropical lower stratosphere, accompanied by an increase in the planetary-wave flux emanating from the troposphere (Fig. 2). Run A confirmed this picture qualitatively, so we conclude that the increased extra-tropical wave-driving did not arise from a trend in the local wave focusing in the stratosphere, but rather from a larger flux of wave activity from the troposphere in the sub-tropics.

A significant non-zero trend in the mass flux into the stratosphere alters the current projections of tropospheric abundances of CFCs given, for example, in the 1998 World Meteorological Organization ozone assessment¹⁹. These projections are based on emission scenarios and on the rate at which the CFCs pass into the stratosphere where they are photolysed; that is, $d\chi/dt = S(t) - f\epsilon\chi$, where χ is the mean tropospheric mass mixing ratio, $S(t)$ the tropospheric source, f the fraction of the underlying mass upwelling into the stratosphere per unit time, and ϵ the efficiency of the stratospheric photodegradation of the CFC. The atmospheric lifetime of a CFC, τ , is then simply $1/(f\epsilon)$. Chlorine loadings in the stratosphere lag those in the troposphere by the average time required for the air to travel from the troposphere to the stratosphere^{1,19}. An increased mass flux into the stratosphere is balanced by an identical increased outward flux, so the changes in tropospheric abundances of CFCs will subsequently be reflected in stratospheric chlorine loadings.

In the lower stratosphere, CFC photolysis rates increase abruptly with height (by orders of magnitude in a few kilometres; see Fig. 4.54 of ref. 20) and the CFCs are then photolysed sufficiently quickly for ϵ to be unaffected by the more rapid transport reducing the average time air spends above the lower stratosphere. Similarly, feedbacks from chemical composition changes resulting from the differences in CFC amounts we estimate below will perturb ϵ —and hence the CFC transient lifetimes—by only a small amount (see Fig. 5.1 of ref. 21) compared to our predicted 3% per decade increase in the mass flux, or f . This increase in f will reduce the

lifetimes of CFCs in such a way that the amounts of CFC11 and CFC12 otherwise expected in 2050 will occur 5 years earlier (Fig. 3). By 2080, this difference increases to 10 years. Moreover, the effect is robust to a variety of future scenarios: for zero future emissions, these speed-up times do not change significantly and, if emissions fall more slowly, the speed-up times increase. Similarly, the speed-up times are largely insensitive to the variations in ϵ that could arise through stratospheric change—for example, if the ozone recovery reduced the influx of photolysing radiation: linear 50% decrease in ϵ from 1990 to 2090 changes our estimated speed-up times by only a few months.

Projected amounts of chlorine in the atmosphere may also have to be revised in future because of unexpected changes in emissions and/or trends in CFC photolysis rates, but we have shown that such revised projections will not affect the speed-up time resulting from faster transport. By the middle of the century, the speed-up in removal of CFCs is sufficiently large to affect the accuracy of ozone predictions in chemistry–climate models that do not explicitly represent transport of CFCs.

Previous studies have concluded that climate change owing to increasing greenhouse-gas concentrations will have a detrimental effect on the ozone layer over the next few decades^{22–25}. Here we have shown that, on the longer timescale, the effects of any increased ozone chemical loss rates will be moderated by a more-rapid-than-projected reduction in the chlorine burden. This results from a predominantly wave-driven strengthening of the meridional transporting circulation which cannot be properly reproduced by two-dimensional models used in projecting future chlorine amounts^{19,21}. Equally, adequate stratospheric resolution is needed to represent the wave dissipation that drives the increase in circulation, and so only those models with a comprehensive stratospheric component²⁶ will be able to properly capture this effect.

We emphasize, however, that while CFC levels remain artificially high, the ozone layer will continue to be vulnerable to the enhanced destruction resulting from a greenhouse cooling of the stratosphere^{22–25}. Also, we note that the accelerated meridional transport could affect the ozone distribution in other ways (for example, by changing the adiabatic warming and cooling, or enhancing the

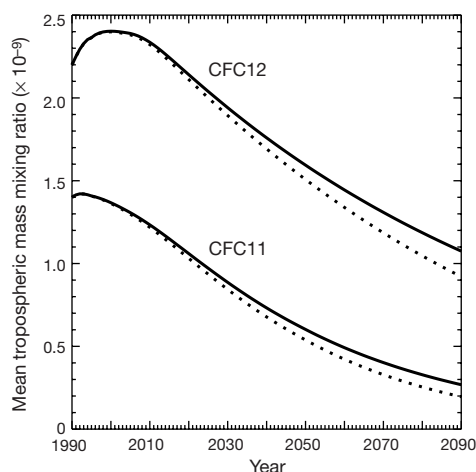


Figure 3 Predicted tropospheric mass mixing ratio for CFC11 and CFC12. The solid curves show the predicted amounts assuming the current specified lifetimes of 45 and 100 years for CFC11 and CFC12, respectively³⁰. The dotted curve shows amounts after the lifetimes have been adjusted to allow for a 3% per decade increase in the upward mass flux at 68 hPa. Initial values and projected sources are taken from Table 11-5 of ref. 19.

transport of ozone from the tropics to the polar regions) though, unlike the CFC removal, these processes are already represented in chemistry–climate models^{22,24,25,27}. Possible changes to the spatial distributions of the various reactive species may also affect details of the ozone response but, ultimately, ozone depletion will recover sooner as a result of the faster removal of CFCs. Finally, the effect described here has further ramifications for climate prediction. This is because the abundance of any greenhouse gas which, like the CFCs, is transported from the troposphere and then destroyed in the stratosphere (for example, CH₄ and N₂O) will likewise² be reduced by an acceleration of the transport from the troposphere into the stratosphere. □

Received 18 September 2000; accepted 9 February 2001.

1. Stratospheric Ozone 1999 (UK Stratospheric Ozone Review Group, Seventh report, HMSO, 1999).
2. Holton, J. R. *et al.* Stratosphere-troposphere exchange. *Rev. Geophys.* **33**, 403–439 (1995).
3. Butchart, N., Austin, J., Knight, J. R., Scaife, A. A. & Gallani, M. L. The response of the stratospheric climate to projected changes in the concentrations of the well-mixed greenhouse gases from 1992–2051. *J. Clim.* **13**, 2142–2159 (2000).
4. Butchart, N. & Austin, J. Middle atmosphere climatologies from the troposphere-stratosphere configuration of the UKMO's Unified Model. *J. Atmos. Sci.* **55**, 2782–2809 (1998).
5. Cullen, M. J. P. The unified forecast/climate model. *Meteorol. Mag.* **122**, 81–94 (1983).
6. Gregory, D., Shutts, G. J. & Mitchell, J. R. A new gravity-wave-drag scheme incorporating anisotropic orography and low-level wave breaking: Impact upon the climate of the UK Meteorological Office Unified Model. *Q. J. R. Meteorol. Soc.* **124**, 463–493 (1998).
7. Mitchell, J. F. B., Johns, T. C., Gregory, J. M. & Tett, S. F. B. Climate response to increasing levels of greenhouse gases and sulphate aerosols. *Nature* **376**, 501–504 (1995).
8. Houghton, J. T. *et al.* (eds) *Climate Change 1995* Ch. 6 (Cambridge Univ. Press, Cambridge, 1996).
9. Rosenlof, K. H. Seasonal cycle of the residual mean meridional circulation in the stratosphere. *J. Geophys. Res.* **100**, 5173–5191 (1995).
10. Mote, P. W. *et al.* An atmospheric tape recorder: The imprint of tropical tropopause temperatures on stratospheric water vapor. *J. Geophys. Res.* **101**, 3989–4006 (1996).
11. Mote, P. W. *et al.* Vertical velocity, vertical diffusion, and dilution by midlatitude air in the tropical lower stratosphere. *J. Geophys. Res.* **103**, 8561–8666 (1998).
12. Yulaeva, E., Holton, J. R. & Wallace, J. M. On the cause of the annual cycle in tropical lower-stratospheric temperatures. *J. Atmos. Sci.* **51**, 169–174 (1994).
13. Andrews, D. G. & McIntyre, M. E. Planetary waves in horizontal and vertical shear: The generalized Eliassen-Palm relation and the mean zonal acceleration. *J. Atmos. Sci.* **33**, 2031–2048 (1976).
14. Andrews, D. G. & McIntyre, M. E. Generalized Eliassen-Palm and Charney-Drazin theorems for waves on axisymmetric mean flows in compressible atmospheres. *J. Atmos. Sci.* **35**, 175–185 (1978).
15. Mahfouf, J. F., Cariolle, D., Royer, J.-F., Gely, J.-F. & Timbal, B. Response of the Météo-France climate model to changes in CO₂ and sea surface temperature. *Clim. Dyn.* **9**, 345–362 (1994).
16. Rind, D., Shindell, D., Loneragan, P. & Balachandran, N. K. Climate change in the middle atmosphere. Part III: the doubled CO₂ climate revisited. *J. Clim.* **11**, 876–894 (1998).
17. Haynes, P. H., Marks, C. J., McIntyre, M. E., Shepherd, T. G. & Shine, K. P. On the “downward control” of the extratropical diabatic circulations by eddy-induced mean zonal forces. *J. Atmos. Sci.* **48**, 651–678 (1991).

18. Rosenlof, K. H. & Holton, J. R. Estimates of the stratospheric residual circulation using the downward control principle. *J. Geophys. Res.* **98**, 10465–10479 (1993).
19. Madronich, S. *et al.* in *Scientific Assessment of Ozone Depletion: 1998* Ch. 11 (World Meteorological Organization, Geneva, 1999).
20. Brasseur, G. & Solomon, S. *Aeronomy of the Middle Atmosphere* (Reidel, Dordrecht, 1984).
21. Ko, M. K. W. *et al.* in *Report on the Concentrations, Lifetimes, and Trends of CFCs, Halons, and Related Species* (eds Kaye, J. A., Penkett, S. A. & Ormond, F. M.) Ch. 5 (Reference publication 1339, NASA, Washington DC, 1994).
22. Shindell, D. T., Rind, D. & Loneragan, P. Increased polar stratospheric ozone losses and delayed eventual recovery owing to increasing greenhouse-gas concentrations. *Nature* **392**, 589–592 (1998).
23. Waibel, A. E. *et al.* Arctic ozone loss due to denitrification. *Science* **283**, 2064–2068 (1999).
24. Austin, J., Knight, J. & Butchart, N. Three-dimensional chemical model simulations of the ozone layer: 1979–2015. *Q. J. R. Meteorol. Soc.* **126**, 1533–1556 (2000).
25. Austin, J., Butchart, N. & Knight, J. Three-dimensional chemical model simulations of the ozone layer: 2015–2055. *Q. J. R. Meteorol. Soc.* (in the press).
26. Pawson, S. *et al.* The GCM-Reality Intercomparison Project for SPARC (GRIPS): Scientific issues and initial results. *Bull. Am. Meteorol. Soc.* **81**, 781–795 (2000).
27. Hofmann, D. J. *et al.* in *Scientific Assessment of Ozone Depletion: 1998* Ch. 12 (World Meteorological Organization, Geneva, 1999).
28. Andrews, D. G., Holton, J. R. & Leovy, C. B. *Middle Atmosphere Dynamics* Ch. 3 (Academic, London, 1987).
29. Dunkerton, T., Hsu, C.-P. F. & McIntyre, M. E. Some Eulerian and Lagrangian diagnostics for a model stratospheric warming. *J. Atmos. Sci.* **38**, 819–843 (1981).
30. Prinn, R. G. *et al.* in *Scientific Assessment of Ozone Depletion: 1998* Ch. 1 (World Meteorological Organization, Geneva, 1999).

Acknowledgements

We thank J. Austin, B. Callander, R. Derwent, J. Knight and R. Swinbank for comments. This work was supported by the Public Meteorological Services Research Programme of The Met Office, and the UK Department of the Environment, Transport and the Regions.

Correspondence and requests for materials should be addressed to N.B. (e-mail: nburchart@meto.gov.uk).

Control of cation concentrations in stream waters by surface soil processes in an Amazonian watershed

Daniel Markewitz^{*†}, Eric A. Davidson^{*}, Ricardo de O. Figueiredo[‡], Reynaldo L. Victoria[§] & Alex V. Krusche[§]

^{*} The Woods Hole Research Center, Woods Hole, Massachusetts 02543, USA

[‡] Instituto de Pesquisa Ambiental da Amazônia (IPAM), Belém, Pará, 66035-100 Brazil

[§] Centro de Energia Nuclear na Agricultura (CENA), Piracicaba, SP, 13400-970 Brazil

The chemical composition of ground waters and stream waters is thought to be determined primarily by weathering of parent rock^{1–5}. In relatively young soils such as those occurring in most temperate ecosystems, dissolution of primary minerals by carbonic acid is the predominant weathering pathway that liberates Ca²⁺, Mg²⁺ and K⁺ and generates alkalinity in the hydrosphere⁶. But control of water chemistry in old and highly weathered soils that have lost reservoirs of primary minerals (a common feature of many tropical soils) is less well understood. Here we present soil and water chemistry data from a 10,000-hectare watershed on highly weathered soil in the Brazilian Amazon. Streamwater cation concentrations and alkalinity are positively correlated to each other and to streamwater discharge, suggesting that cations

[†] Present address: Warnell School of Forest Resources, University of Georgia, Athens, Georgia 30602, USA.

and bicarbonate are mainly flushed from surface soil layers by rainfall rather than being the products of deep soil weathering carried by groundwater flow. These patterns contrast with the seasonal patterns widely recognized in temperate ecosystems with less strongly weathered soils^{2,7}. In this particular watershed, partial forest clearing and burning 30 years previously enriched the soils in cations and so may have increased the observed wet season leaching of cations. Nevertheless, annual inputs and outputs of cations from the watershed are low and nearly balanced, and thus soil cations from forest burning will remain available for forest regrowth over the next few decades. Our observations suggest that increased root and microbial respiration during the wet season generates CO₂ that drives cation-bicarbonate leaching, resulting in a biologically mediated process of surface soil exchange controlling the streamwater inputs of cations and alkalinity from these highly weathered soils.

Streamwater cation concentration and stream discharge are usually negatively correlated. During low, base-flow conditions fed primarily by ground water, Ca²⁺, Mg²⁺ and K⁺ are largely derived from weathering of primary minerals, and are typically at their highest concentrations. As stream discharge increases, streamwater solute concentrations decrease because the stream becomes diluted by surface or lateral flow waters that enter during rain events or snowmelt^{1,2,7}. Surface flow waters often have short residence times within the soil and, therefore, have low ionic strengths. But these relationships between solute concentration and stream discharge have been defined in temperate latitudes, and few data exist from tropical landscapes, where many tropical soils on stable geological surfaces have weathered to great depths⁸ and produce extremely dilute solutions in drainage water^{9,10}.

We measured streamwater solution concentrations and discharge over two annual cycles in a watershed of the eastern Amazon. The soils of this region are highly weathered Oxisols, which have formed in Upper Tertiary fluvial-lacustrine sediments common along the Amazon river⁹. Oxisols cover about 45% of the entire Amazonian basin¹¹ although they may form from a diversity of geologic substrates, including the sediments of the Amazon trough and the rocks of the Brazilian and Guyana shields. These soils generally contribute dilute waters (total cation charge < 400 $\mu\text{equiv. l}^{-1}$) to the main stem of the Amazon river⁹.

The current landscape of the study watershed is a mosaic of primary (29%) and secondary forests (23%), and managed (12%) and degraded (23%) pastures. Remnant forests remain in hilltop positions, and cattle pastures typically extend very near to the edge of the first-order stream. The solution chemistry of the stream, Igarapé 54, draining this watershed is an integrated expression of the biogeochemistry in these diverse ecosystems.

The chemical concentrations of individual elements in Igarapé 54 fluctuated in response to the seasonal patterns of rainfall and stream discharge (Fig. 1). The solutions draining this landscape are dilute all year round, with mean conductivities < 40 $\mu\text{S cm}^{-1}$ and a range of 25–45 $\mu\text{S cm}^{-1}$ (Fig. 1). Despite this low electrolyte concentration, both pH and alkalinity were highly responsive to annual cycles of stream discharge (Fig. 1). The pH of stream water ranged from 4.5 to 6.0, from the dry to wet season, respectively, representing more than an order of magnitude difference in the hydrogen ion activity of stream solutions. Similarly, streamwater alkalinity in the dry and wet season ranged from 0 to 200 $\mu\text{equiv. l}^{-1}$ (Fig. 1). These patterns in stream chemistry reveal a shift from an acidic (pH 4.5), dilute (28 $\mu\text{S cm}^{-1}$), low-alkalinity (0–5 $\mu\text{equiv. l}^{-1}$) solution during the base-flow conditions of the dry season, to a solution that during the high-flow conditions of the wet season is approaching neutrality (pH 6), has an increased concentration of dissolved solids (35 $\mu\text{S cm}^{-1}$), and has a substantial increase in acid neutralizing capacity (100–200 $\mu\text{equiv. l}^{-1}$). What is most notable, however, is that these patterns are contrary to those observed in temperate ecosystems, where streamwater alkalinity decreases during periods

of high flow^{1,2,7}.

Like the observed decreases in H⁺ (increasing pH) during periods of high flow, the concentrations of Si (as H₄SiO₄), Na⁺ and Cl[−] also decreased (Fig. 1). In contrast, the concentrations of Ca²⁺, Mg²⁺, K⁺, NO₃[−] and SO₄^{2−} increased with increasing flow, following the same seasonal pattern as HCO₃[−] alkalinity (Fig. 1). The concentration of HCO₃[−] was strongly correlated with the concentration of Ca²⁺ ($r = 0.93$), Mg²⁺ ($r = 0.92$) and K⁺ ($r = 0.79$). Concentrations of dissolved organic carbon, which can also be an important anion, were low (< 1.3 mg l^{−1}), showed little seasonality, and were poorly correlated with any cation. Thus, on a nearly stoichiometric basis, HCO₃[−] alone can explain most of the variability in cation concentrations during high flow. These cation–HCO₃[−] relationships have been commonly observed during base-flow conditions for many streams^{1,2,7} but are remarkable in this case, because they appear to control streamwater chemistry during periods of high flow during the wet season. Current geochemical models, based primarily on observations from temperate regions, generate the preponderance of streamwater HCO₃[−] alkalinity through processes of mineral

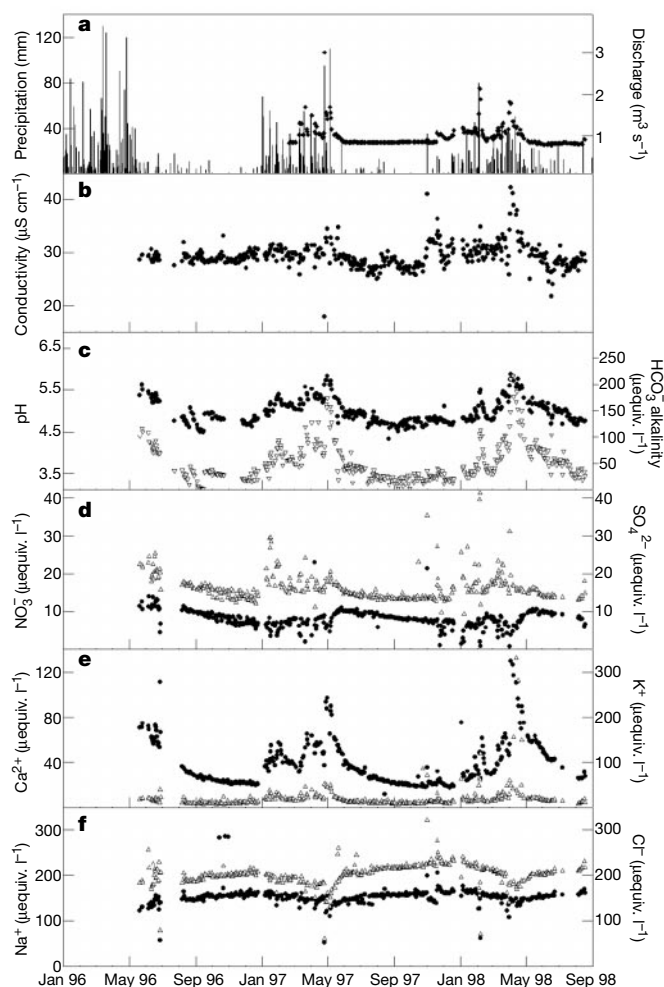


Figure 1 Seasonal hydrologic fluxes and streamwater chemistry from an Amazonian watershed on highly weathered soil. These data illustrate a unique pattern of increased bicarbonate alkalinity and cation concentrations during periods of highest flow. In **a**, precipitation inputs per day are indicated by the vertical bars in relation to the left-hand axis while discharge is shown by the diamonds related to the right-hand axis. In the lower five panels (**b–f**), circles correspond to numbers on the left-hand axis and triangles correspond to numbers on the right-hand axis. We note the seasonal covariation of pH (**c**), bicarbonate alkalinity (**d**), cation concentrations (**e** and **f**), and discharge (**a**), and the lack of strong covariation with nitrate (**d**) and sulphate (**d**).

weathering in bedrock, with these relationships predominating during low-flow conditions^{7,12}.

The gas–liquid–solid system of CO₂, H₂CO₃ and proton consumption during mineral weathering is the predominant mechanism generating alkalinity in most natural waters¹². The temporal interaction of H₂CO₃ with bedrock weathering is currently treated as a primary driver in the long-term carbon cycle^{13–15}. The HCO₃[−] alkalinity in the stream, surface, and lateral flow waters of this tropical landscape is also produced by the dissociation of dissolved CO₂ (that is, H₂CO₃) generated by root and microbial respiration in the soil. Mean monthly emissions of CO₂ from the soil surface vary by a factor of 5–6 between wet and dry seasons¹⁶, and soil CO₂ can reach concentrations as high as 8% by volume during the rainy season¹⁷. When H₂CO₃ dissociates to H⁺ and HCO₃[−], proton consumption occurs through cation exchange processes on soil surfaces, with subsequent leaching of the cation with HCO₃[−] in solution. In contrast to alkalinity-generating processes in temperate regions, a much greater annual flux of alkalinity is generated in the surface soil than in the underlying groundwater aquifer of this tropical landscape.

The ratio of Na/(Na + Ca) has been used to infer whether stream-water cations are derived from precipitation inputs or mineral weathering². A ratio approaching one indicates little contribution from mineral weathering, while a smaller ratio indicates greater weathering inputs. In this tropical watershed, however, the Na/(Na + Ca) ratio approaches one during the period of lowest flow, when inputs from bedrock weathering in association with ground water are assumed to be greatest (Fig. 2). Conversely, during high flow the ratio is smallest, indicating weathering inputs. In this unglaciated, deep, and highly weathered soil, <0.5% of the soil throughout the upper 5 m is comprised of primary minerals that contain weatherable Ca²⁺, Mg²⁺ and K⁺. These primary minerals are also probably negligible to much greater depths (Table 1). Hence, inputs of Ca²⁺, Mg²⁺ and K⁺ cannot be from weathering of primary minerals, but rather must be from removal of cations from the cation-exchange complex, which are in greatest concentration in surface soils (Table 1).

The sources of cations and of carbonic acid production need to be understood in order to model biogeochemical processes in these tropical landscapes. The impoverished geological sediments and soils of this region may represent a geochemical end member that has not been well studied at the watershed scale. Although an end member, these geochemical conditions may be characteristic of the large fraction of tropical forest ecosystems that are found on

similarly highly weathered deep substrates, where inputs of nutrients from mineral weathering are virtually nil and where most biogeochemical element cycles are controlled through organic matter cycling. Such organically controlled biogeochemical patterns for stream water are well described for nitrogen (which is not present in rocks and thus is predominantly cycled through organic matter in surface soils), but this strong biogeochemical influence of surface soil processes on stream water has not been previously observed for elements derived from primary minerals, such as Ca²⁺, Mg²⁺ or K⁺.

The watershed that we studied has been cleared of about 70% of its original native forest during the past 30 years, with forest remnants remaining on upland areas and pastures being nearer the stream. Conversion of tropical forests to cattle pastures results in export of carbon and plant nutrients from the ecosystem as harvested logs and as emissions from fire^{18,19}. A net hydrologic

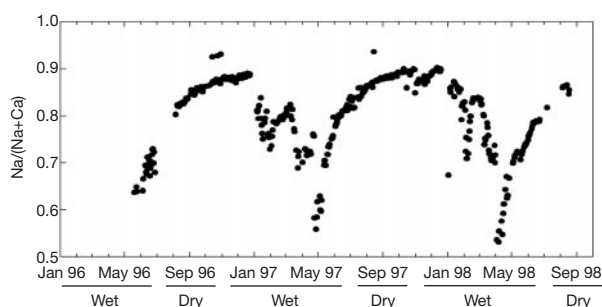


Figure 2 The mineral weathering index Na/(Na + Ca) for stream waters of this Amazonian watershed indicate an anomalous pattern. The ratio approaches one at low stream flow, when weathering inputs of Ca from groundwater sources should be greatest. The ratio of Na/(Na + Ca) is used to infer the contribution of mineral weathering to streamwater cation concentrations; ratios near one indicate low rates of mineral weathering inputs. In temperate ecosystems, where most mineral–solute relationships have been defined, this ratio approaches one during periods of high flow when surface waters are contributing to stream flow. The opposite pattern is shown for this first-order tropical stream draining a highly weathered and deep Oxisol, indicating that Ca is coming predominantly from surface or lateral-flow waters during high flow as a result of leaching of surface soils, and that Ca inputs during groundwater base-flow conditions are minimal. Wet and dry indicate rainfall season and correspond to periods of high flow and base flow, respectively.

Table 1 Soil chemical and physical parameters

Land use	Depth (m)	Bulk density (g cm ^{−3})	pH _s [*]	K [†] (cmol _c kg ^{−1})	Ca (cmol _c kg ^{−1})	Mg (cmol _c kg ^{−1})	ECEC [‡] (cmol _c kg ^{−1})
Primary forest	0–0.1	1.02 (3)a	3.77 (3)a	0.085 (2)a	1.252 (36)a	0.606 (17)a	2.80 (8)a
	0.1–0.2	1.18 (2)	3.72 (9)	0.043 (2)	0.398 (65)	0.311 (48)	1.93 (5)
	0.2–0.5	1.26 (3)	3.96 (2)	0.023 (92)	0.200 (35)	0.251 (59)	1.39 (13)
	0.5–1	1.28 (2)	4.14 (3)	0.006 (98)	0.065 (47)	0.161 (52)	0.77 (21)
	1–2	1.31 (1)	4.27 (2)	0.004 (105)	0.045 (52)	0.209 (55)	0.56 (12)
	2–3	1.32 (1)	4.32 (2)	0.003 (77)	0.010 (137)	0.137 (74)	0.32 (29)
	3–4	1.31 (2)	4.38 (2)	0.003 (44)	0.001 (185)	0.065 (76)	0.18 (51)
	4–5	1.31 (2)	4.43 (3)	0.003 (59)	0.000	0.037 (87)	0.15 (67)
	5–6	1.26 (10)	4.36 (3)	0.004 (28)	0.002 (283)	0.025 (85)	0.17 (48)
	7–7.5	1.30 (3)	4.42 (3)	0.003 (84)	0.001 (224)	0.015 (91)	0.11 (42)
Secondary forest	0–0.1	1.01 (10)a	4.56 (3)b	0.12 (17)a	3.43 (18)b	1.18 (19)a	4.88 (17)b
	0.1–0.2	1.14 (4)	4.14 (2)	0.06 (17)	1.43 (31)	0.64 (11)	2.50 (15)
Degraded pasture	0–0.1	1.22 (1)b	4.55 (8)b	0.26 (38)b	2.68 (46)ab	1.00 (41)a	4.25 (33)ab
	0.1–0.2	1.23 (3)	3.90 (4)	0.08 (25)	0.57 (51)	0.33 (58)	1.98 (30)
Managed pasture	0–0.1	1.24 (1)b	4.70 (2)b	0.19 (26)a	2.69 (12)ab	0.87 (8)a	3.86 (9)ab
	0.1–0.2	1.24 (3)	4.55 (3)	0.12 (50)	1.45 (12)	0.46 (15)	2.20 (13)

Data for surface soils from contrasting land uses in Paragominas, Brazil, showing mean (coefficient of variation). In the depth range 0–0.1 m, letters are results of post hoc tests between land uses using Tukey's honestly significant difference for analysis of variance with $P < 0.05$ (except ECEC, $P < 0.08$). No important differences in soil chemistry were observed among land uses at depths below 0.2 m.

^{*}Soil pH measured in a 0.01 M CaCl₂ salt solution in a 1:2 w/v ratio.

[†]Exchangeable cations were estimated by extraction with Mehlich I solution in a 1:10 w/v ratio.

[‡]ECEC is effective cation exchange capacity by sum of cations method.

export of Ca^{2+} , Mg^{2+} and K^+ may also be occurring. Surface soils in cattle pastures and secondary forests are enriched in cations derived from the forest plant biomass that were input to the soil as ash during forest clearing and burning (Table 1). If biogenic H_2CO_3 is gradually re-acidifying these surface soils²⁰ with the subsequent leaching of cations and HCO_3^- to the stream, the magnitude of observed seasonal peaks of streamwater concentrations could be at least partly the result of these land use changes. Our best estimates of the annual hydrologic export of cations in the stream, however, are not significantly different from our estimated rainfall inputs of 3, 1 and $5 \text{ kg ha}^{-1} \text{ yr}^{-1}$ for Ca^{2+} , Mg^{2+} and K^+ , respectively. Hence any input–output imbalance is smaller than our detection accuracy (probably $< 1 \text{ kg ha}^{-1} \text{ yr}^{-1}$). The relatively large seasonal variation in streamwater cation concentrations does not necessarily indicate a large net hydrologic export. The process of re-acidification of pasture soils—after forest ash inputs of approximately 300, 100 and 250 kg ha^{-1} of respectively Ca^{2+} , Mg^{2+} and K^+ —will require many decades. Such a slow rate of cation loss means that cations could remain available for forest re-growth during the next few decades.

We do not know if the wet season hydrologic export of cations also occurs in watersheds of mostly undisturbed mature forests on the same soil substrate, which are rare in the eastern Amazon region. It is possible that forest clearing and burning may simply be enhancing a biogeochemical process that was already present, with a similar, but perhaps less pronounced, seasonal pattern. Alternatively, the pre-cleared condition may have shown uniformly low cation concentrations, alkalinity and pH throughout the year. In any case, the inputs of cations and HCO_3^- to the stream water from either forest or pasture soils of this tropical watershed are not controlled by the geochemical relationships of the lower soil profile, but rather are controlled by the biogeochemical cycling within the plant/surface-soil component of the ecosystem. In this tropical ecosystem, even the elements originally derived from primary minerals are now predominantly controlled by organic matter cycles and the charge properties of surface soil. A biologically mediated, near-surface process has replaced deep soil mineral weathering as the process controlling the contributions of cations and bicarbonate alkalinity to stream water. Identifying the sources and mechanisms of solute contribution to Amazonian streams is necessary for understanding nutrient cycling processes in mature tropical forests and the long-term effects of deforestation²¹ occurring in the region. □

Methods

Research area

The Fazenda Vitória is an active cattle ranch 6 km north of the town of Paragominas ($2^\circ 59' \text{ S}$, $47^\circ 31' \text{ W}$) in the northern Brazilian state of Pará, with a 15-yr research history^{22,23}. This area was settled in the mid-1960s following the construction of the Belém–Brasília highway. The areas that are now in pasture and secondary forest were originally cleared and burned in 1969. This region of Brazil experiences strong seasonality in rainfall distribution, with $< 15\%$ of the total annual precipitation (1972–94 mean $\pm 1 \text{ s.d.}$, $1,803 \pm 555 \text{ mm}$)²³ falling between June and November. Soils of this region are derived from the Belterra clay formation that developed on the top of a Pleistocene terrace, which is widespread at elevations below 200 m in the Amazon basin. Belterra clays consist mainly of kaolinite, with minor fractions of quartz and haematite. Mineralogical analyses of soils from our site were conducted by O. Chadwick (Univ. California, Santa Barbara). These soils have high clay contents, $> 50\%$, and low effective cation exchange capacity, $< 5 \text{ cmol kg}^{-1}$, throughout the upper 10 m. These deep and highly weathered soils are classified as Latossolos Amarelo distrófico (Haplustox, according to USDA soil taxonomy)²⁴ and are most common in eastern and southern Amazonia where deforestation is concentrated¹¹. Igarapé 54 is a moderately sized ($0.8 \text{ m}^3 \text{ s}^{-1}$ discharge during base flow), perennial, first-order stream that dissects a large portion of Fazenda Vitória.

Streamwater chemistry

Streamwater grab samples were collected in acid-washed, 250-ml polypropylene bottles between 07:00 and 08:00 on all days a worker was available. Over the course of the study, collections were made on $\sim 55\%$ of all days. Bottles were filled to capacity to minimize headspace, and were placed in cold storage within a few hours of collection. Approximately 10% of samples were analysed the same day of collection, but most were retained

cold for up to 1 week until analysis. Stream gauge height was measured on all days of collection, and stream discharge was estimated from a gauge height versus discharge relationship established for the stream²⁵. Only six height measurements exceeded the rating curve, which proved sufficient to estimate discharge for $> 98\%$ of all samples. Samples were analysed in a field laboratory in Paragominas—for conductivity, pH and alkalinity—and were then filtered through $0.4\text{-}\mu\text{m}$ polycarbonate filters and kept cold. Alkalinity determinations were estimated by endpoint titrations with 1 mM HCl to $\text{pH } 4.5$ ²⁶. Endpoint titrations can overestimate alkalinity under some conditions, but in a comparison of a subset of 25 samples, the absolute differences from Gran function plots and endpoint titrations to $\text{pH } 4.5$ averaged $16 \mu\text{equiv. l}^{-1}$. Differences between titration to $\text{pH } 5.0$ and 4.5 ($n = 380$) were also small, always being $< 40 \mu\text{equiv. l}^{-1}$. Endpoint titrations were chosen because this more rapid method permitted timely analyses of a larger number of samples. Cation–anion balances of stream solutions averaged 0.95 with a 95% confidence interval of $0.94\text{--}0.96$ ($n = 414$), which also indicates that overestimation of alkalinity was probably minimal. Measurements of Cl^- , SO_4^{2-} and NO_3^- were by ion chromatography, and Ca^{2+} , Mg^{2+} , K^+ and Na^+ were measured by atomic absorption spectrophotometry after addition of $> 10,000 \text{ mg l}^{-1} \text{ LaCl}_3$. ERA (Arvada, Colorado) quality control standards were used to maintain quality assurance over time.

Received 9 June 2000; accepted 2 February 2001.

- Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* 128–175 (Princeton Univ. Press, Princeton, 1984).
- Drever, J. I. *The Geochemistry of Natural Waters; Surface and Groundwater Environments* 3rd edn 138–196 (Prentice Hall, Upper Saddle River, 1997).
- Lasaga, A. C., Soler, J. M., Ganor, J., Burch, T. E. & Nagy, K. L. Chemical weathering rate laws and global geochemical cycles. *Geochim. Cosmochim. Acta* **58**, 2361–2386 (1994).
- Garrels, R. M. & Mackenzie, F. T. in *Equilibrium Concepts in Natural Water Systems* (ed. Stumm, W.) 222–242 (American Chemical Society, Washington DC, 1967).
- Meybeck, M. Global chemical weathering of surficial rocks estimated from river dissolved loads. *Am. J. Sci.* **287**, 401–428 (1987).
- Stumm, W. & Morgan, J. J. *Aquatic Chemistry: Chemical Equilibria and Rates in Natural Waters* 3rd edn 872–932 (Wiley, New York, 1996).
- Likens, G. E. & Bormann, F. H. *Biogeochemistry of a Forested Ecosystem* 2nd edn 46–101 (Springer, New York, 1995).
- Sombroek, W. G. in *The Amazon: Limnology and Landscape Ecology of a Mighty Tropical River and its Basin* (ed. Sioli, H.) 521–536 (Junk, Boston, 1984).
- Stallard, R. F. & Edmond, J. M. Geochemistry of the Amazon II. The influence of geology and weathering environment on the dissolved load. *J. Geophys. Res.* **88**, 9671–9688 (1983).
- Furch, K. in *The Amazon: Limnology and Landscape Ecology of a Mighty Tropical River and its Basin* (ed. Sioli, H.) 167–200 (Junk, Boston, 1984).
- Richter, D. D. & Babbler, L. I. Soil diversity in the tropics. *Adv. Ecol. Res.* **21**, 315–389 (1991).
- Ohte, N. & Tokuchi, N. Geographical variation of the acid buffering of vegetated catchments: Factors determining the bicarbonate leaching. *Glob. Biogeochem. Cycles* **13**, 969–996 (1999).
- Berner, R. A. The rise of plants and their effect on weathering and atmospheric CO_2 . *Science* **276**, 544–545 (1997).
- Berner, R. A. Weathering, plants, and the long-term carbon cycle. *Geochim. Cosmochim. Acta* **56**, 3225–3231 (1992).
- Broecker, W. S. & Sanyal, A. Does atmospheric CO_2 police the rate of chemical weathering? *Glob. Biogeochem. Cycles* **12**, 403–408 (1998).
- Davidson, E. A., Verchot, L. V., Cattaneo, J. H., Ackerman, I. L. & Carvalho, J. E. M. Effects of soil water content on soil respiration in forests and cattle pastures of eastern Amazonia. *Biogeochemistry* **48**, 53–69 (2000).
- Davidson, E. A. & Trumbore, S. E. Gas diffusivity and production of CO_2 in deep soils of the eastern Amazon. *Tellus B* **47**, 550–565 (1995).
- Kauffman, J. B., Cummings, D. L., Ward, D. E. & Babbitt, R. Fire in the Brazilian Amazon I: biomass, nutrient pools, and losses in slashed primary forests. *Oecologia* **104**, 397–408 (1995).
- Ewel, J., Berish, C., Brown, B., Price, N. & Raich, J. Slash and burn impacts on a Costa Rican wet forest site. *Ecology* **62**, 816–829 (1981).
- McColl, J. G. & Cole, D. W. A mechanism of cation transport in a forest soil. *Northwest Sci.* **42**, 134–40 (1968).
- Houghton, R. A. *et al.* Annual fluxes of carbon from deforestation and regrowth in the Brazilian Amazon. *Nature* **403**, 301–304 (2000).
- Uhl, C., Buschbacher, R. & Serrao, E. A. S. Abandoned pastures in eastern Amazonia. I. Patterns of plant succession. *J. Ecol.* **76**, 663–681 (1988).
- Nepstad, D. C. *et al.* The deep-soil link between water and carbon cycles of Amazonian forests and pastures. *Nature* **372**, 666–669 (1994).
- Sombroek, W. G. *Amazon Soils, a Reconnaissance of the Soils of the Brazilian Amazon Region* (PUDOC, Wageningen, Netherlands, 1996).
- Rantz, S. E. *Measurement and Computation of Stream Flow* Vol. 1, Measurement of Stage Height and Discharge (USGS Water Supply Paper 2175, US Government Printing Office, Washington DC, 1982).
- Clesceri, L. S. *et al.* (eds) *Standard Methods for the Examination of Water and Wastewater* 20th edn 2–26–2–29 (United Book Press, Baltimore, 1998).

Acknowledgements

We thank P. Heine, P. Moutinho, I. d. N. Soares, D. C. Nepstad, O. Qafoku, J. A. Raikes, D. D. Richter, L. Verchot, and the staff at the Instituto de Pesquisa Ambiental da Amazônia and the Woods Hole Research Center for analytical and logistical support. This work was supported in part by the A. W. Mellon Foundation, the Environmental Geochemistry Biogeochemistry program of the NSF, the Woods Hole Research Center and the University of Georgia.

Correspondence and requests for materials should be addressed to A.B. (e-mail: dmarke@smokekey.forestry.uga.edu).

Plateau 'pop-up' in the great 1897 Assam earthquake

Roger Bilham*† & Philip England†

* CIRES & Geological Sciences, University of Colorado, Boulder, Colorado 80309-0399, USA

† Earth Sciences, Oxford University, Parks Road, Oxford OX1 3PR, UK

The great Assam earthquake of 12 June 1897 reduced to rubble all masonry buildings within a region of northeastern India roughly the size of England, and was felt over an area exceeding that of the great 1755 Lisbon earthquake¹. Hitherto it was believed that rupture occurred on a north-dipping Himalayan thrust fault propagating south of Bhutan²⁻⁵. But here we show that the northern edge of the Shillong plateau rose violently by at least 11 m during the Assam earthquake, and that this was due to the rupture of a buried reverse fault approximately 110 km in length and dipping steeply away from the Himalaya. The stress drop implied by the rupture geometry and the prodigious fault slip of 18 ± 7 m explains epicentral accelerations observed to exceed 1g vertically and surface velocities exceeding 3 m s^{-1} (ref. 1). This quantitative observation of active deformation of a 'pop-up' structure confirms that faults bounding such structures can penetrate the whole crust. Plateau uplift in the past 2–5 million years has caused the Indian plate to contract locally by $4 \pm 2 \text{ mm yr}^{-1}$, reducing seismic risk in Bhutan but increasing the risk in northern Bangladesh.

Central to our analysis are the long-neglected survey reports of Bond⁶ who located and remeasured the original points of the 1862 trigonometrical survey across the Shillong plateau. Despite the discovery, under appalling conditions, of 8 m of uplift and 4 m of displacement of parts of this plateau, his results were dismissed⁷, because they failed to meet the triangle closure standards of the

Survey of India. Bond, and later Oldham¹, who attempted to interpret the data, suspected that closure errors were caused by continuing movements following the mainshock, an idea that was many years ahead of its time. Oldham, noting the existence of a northward increasing strain gradient in Bond's data, recommended remeasurement of a survey along the northern edge of the plateau, but this was not to occur until 1936⁸ (Fig. 1). These northern measurements had a mean error typical of Survey of India accuracies ($3.3 \mu\text{rad}$), but did not overlap the earlier re-survey and, as in Bond's survey, included neither scale nor azimuth constraints.

The results of the surveys are available as the locations of points^{7,9} with their apparent post-seismic displacements, calculated holding fixed two points within each network^{1,6,8,10}. We seek the values of these parameters that best fit the observed angle changes. The absence of scale and orientation information permits only the analysis of angular changes^{11,10}, which we derive from these published data. We treat the angular changes as though they reflect deformation of an elastic half-space, distorted by slip on a single rectangular plane representing the 1897 earthquake. In doing this, we neglect deformation caused by other earthquakes between 1860–69 and 1897 and by post-seismic deformation before the re-surveys.

The surface distortion caused by slip on a buried dislocation can be calculated from nine parameters that describe its geometry and slip¹². The problem is sufficiently nonlinear that many local minima exist; we therefore sought a global minimum by systematically searching parameter space. Results of this search are contoured in Fig. 2b. The best-fitting solution, in the sense of minimizing the misfits to the angular changes normalized by their uncertainties, corresponds to a slip of 16 m on a fault plane striking ESE for 110 km and dipping SSW at 57° beneath the northern edge of the plateau (Fig. 2); slip on the plane extends from 9 to 45 km beneath the surface, with a rake of 76° .

Several features of the solution are unexpected. First, all previous studies favoured slip on a plane dipping northwards from the south of the plateau. Furthermore, the extension of the rupture plane to

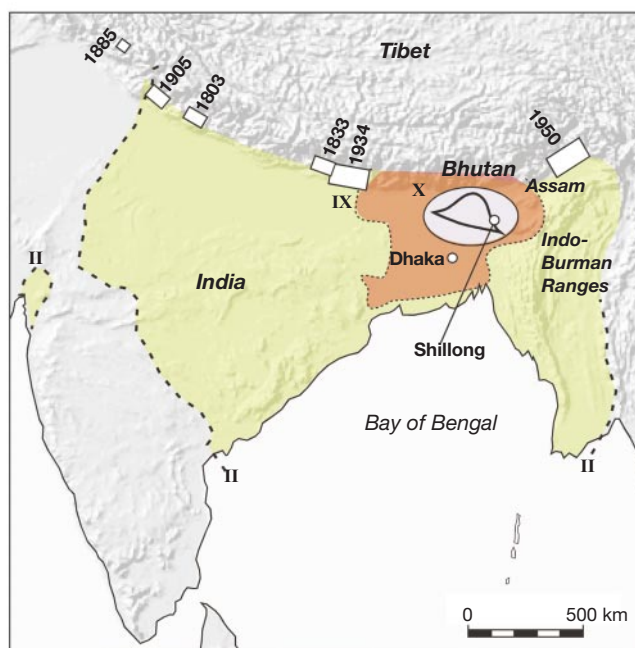


Figure 1 Area shaken by the great 1897 Assam earthquake, and location of major Himalayan ruptures in the past 200 years³⁰. Masonry structures were damaged within Oldham's¹ intensity IX contour (orange), and destroyed within the intensity X ellipse (violet). The earthquake was felt by persons within the intensity II region (green). The

orientation of the curious 'Mexican-hat' shape of the epicentral region mapped by Oldham corresponds to the strike of the causal subsurface rupture derived from geodetic data (Fig. 2).

the surface corresponds to no mapped fault¹³. In the absence of a named fault we refer to rupture occurring on the 'Oldham fault'. The easternmost edge of the Oldham fault, and the latitude of its western edge, are well constrained by the triangulation data (Fig. 2b). The longitude of the westward termination is consistent with 10 m of up-to-the east vertical slip on the Chedrang fault (Fig. 2). Boundary element¹⁴ calculations treating the Chedrang fault as a frictionless surface driven by coseismic stress changes caused by slip on the Oldham fault emulate the observed distribution of slip only when the ends of the two faults approach within a few kilometres. Finally, the projection of the preferred rupture plane to the surface follows slope breaks on an ESE topographic

escarpment along the northern edge of the plateau (Fig. 2c); this projection also follows the strike and location of Oldham's 'Mexican-hat' outline of the rupture zone¹ corresponding to the zone of highest shaking intensity and aftershock productivity¹ (Fig. 1).

Our best-fitting value of 16 m for the slip is one of the greatest for any measured earthquake. A lower bound on the magnitude of this slip comes from the magnitude of the shear strains near the top of the rupture plane (Fig. 2) and, independently, from Oldham's observations on the Chedrang fault, which imply a minimum of 11 m of displacement on the main fault plane. There is a trade-off between the depth of the top of the rupture and the magnitude of

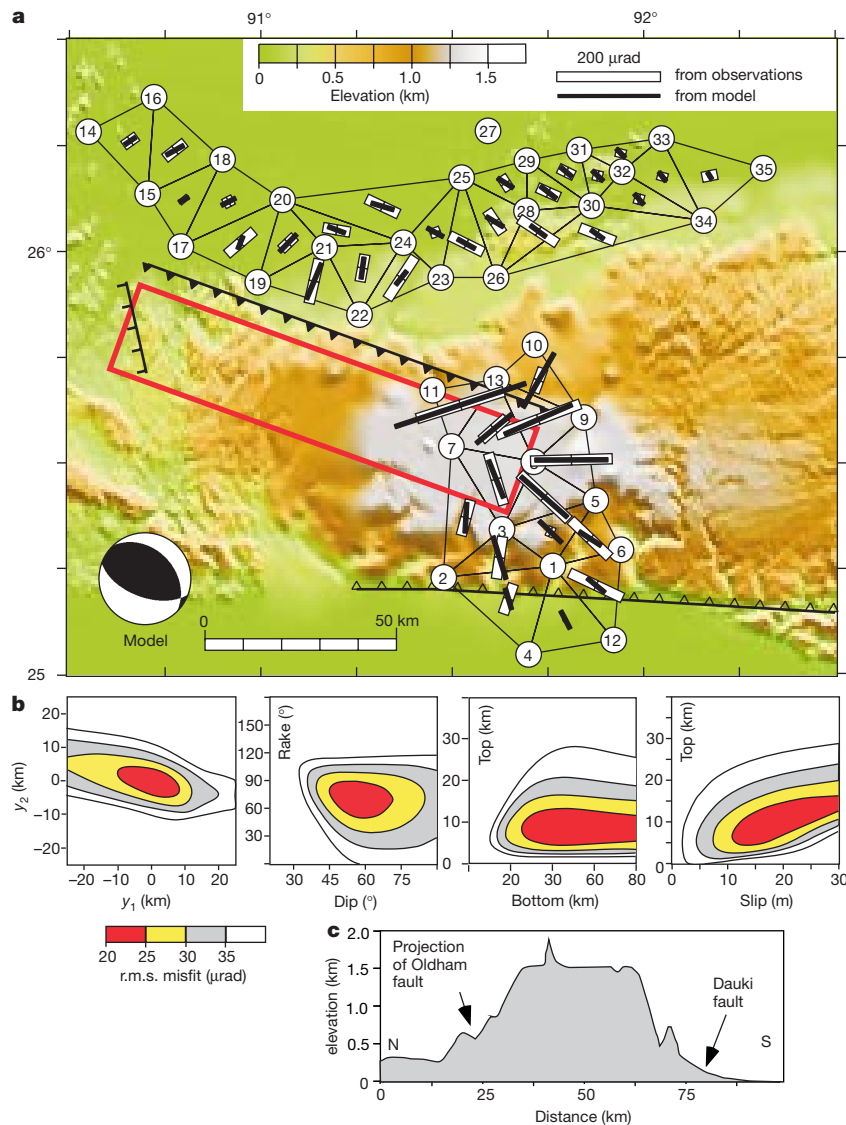


Figure 2 Epicentral region of the Great 1897 Assam earthquake showing geodetic data and inferred rupture plane. **a**, Trigonometrical stations remeasured on⁶ and north⁸ of the Shillong plateau following the 1897 earthquake. Shear strains for each triangle are displayed as bars whose orientations are parallel to the smallest (most compressional) principal strain and whose lengths are proportional to the shear strain, $\Gamma = (\gamma_1^2 + \gamma_2^2)^{1/2}$ where $\gamma_1 = \delta u / \delta x - \delta v / \delta y$, $\gamma_2 = \delta u / \delta y + \delta v / \delta x$, and u , v are components of displacement in the easterly (x) and northerly (y) directions¹¹. White rectangles are calculated for the triangulation observations, and black bars show the strains calculated for the best-fitting planar dislocation. Red rectangle indicates subsurface location of this southwest dipping dislocation; thick black line with filled triangles shows the surface intersection of the continuation of this plane to the land surface (slip terminated 9 km

below the surface). Short black line at western edge of fault plane indicates location of Chedrang fault, where Oldham¹ observed 11 m of co-seismic slip down to the west. Line with open triangles to south of the plateau shows location of the Dauki fault. **b**, Trade-offs between parameters of the model dislocation. Each panel displays the variation in root-mean-square misfit to the observed angular changes, with all parameters except those illustrated held fixed to their best-fitting values; y_1 and y_2 refer to latitudinal coordinate of, respectively, the western and eastern ends of the fault. **c**, Topographic section through the central Shillong plateau showing projected surface location of the Oldham fault. A peneplanation surface is evident at ~ 1.6 km. One of several isolated peaks that rise to ~ 2 km is shown on this section.

the slip (Fig. 2a), with greater slip (up to 25 m) allowed if the fault is buried more deeply. Slip of less than 10 m, however, produces significantly worse fits to the observations. The dip of the fault and the rake of the slip are constrained to within 15° by the distribution of shear strains (Fig. 2a); specifically, a dip of less than 45° would produce strains in the northern network that are excluded by the data. The NNE–SSW compressional strain implied by our rupture is consistent with the P axes of smaller earthquakes in the region during this century^{15,16,17}.

The maximum depth of the lower edge of the rupture is not well constrained by the observations (Fig. 2a). The geodetic data suggest that rupture extended to a depth of at least 35 km, and may well have cut through to the base of the crust (Fig. 3), here estimated to lie at 43–46 km depth¹⁸. This suggestion is consistent with the evidence that earthquakes occur to depths of 30–50 km beneath the Shillong plateau¹⁵. The unusually high ratio of slip to fault length implies a static stress drop at the high end of the observed range, consistent with the violence of the event. It was in this earthquake that, for the first time, accelerations exceeding 1g were identified as responsible for propelling objects into the air. From European seismograms, Richter calculated a surface-wave magnitude of $M_s = 8.7$ although retrospective calibration of these same records yields $M_s = 8.0 \pm 0.1$ (ref. 19). The parameters of the rupture shown in Fig. 2 correspond to moment magnitude $M_w = 8.1$.

The triangulation data exclude significant slip on the Dauki fault, at the southern margin of the plateau, during the 1897 event. The question remains as to whether this undoubtedly major fault^{20,21} is a Himalayan thrust fault, as previous interpretations of the 1897 event have concluded. In Miocene times this fault accommodated westward translation of the Indo–Burman ranges²², but, increasingly since Pliocene times, the fault has permitted reverse slip. If the Dauki fault is, indeed, a major fault bounding the plateau, then it is reasonable to assume that, like the Oldham fault, it cuts to the base of the crust. If the Dauki fault were a gently dipping thrust fault (6° – 15° N) it would intersect the Oldham fault in the mid-crust (6–14 km) (Fig. 3). For the two faults not to intersect within the crust, the Dauki fault would need to dip at more than 40° .

We therefore interpret the Shillong plateau as a pop-up structure bounded by (at least) two reverse faults (Fig. 3). This interpretation is consistent with, but not required by, gravity data that suggest the plateau is uncompensated^{15,18}. It is unclear why the Shillong plateau should be the only major pop-up structure in the northern Indian shield. The in-plane compressional stress is principally responsible for driving the uplift of the plateau, yet the associated flexure due to the bending of the Indian plate may contribute to the initiation and/or sustenance of plateau pop-up. Moreover, two major loads on the lithosphere are present here that are absent elsewhere along the Himalayan chain: the sediments of the Bengal fan are up to 23 km thick to the south and east of the plateau (Fig. 3), and

the Indo–Burman ranges represent an overthrust load that is moving westwards over the Indian lithosphere²². Either, or both, of these loads may amplify the stresses already acting on the Indian plate due to the load of the Himalaya. Alternatively, the explanation may lie in a pre-existing heterogeneity; the Miocene strike-slip Dauki fault may have nucleated dip-slip faulting as the region approached the Himalaya in Pliocene times.

We now estimate the rate of uplift and shortening of the Shillong plateau. The time of uplift of this plateau has been inferred from a coarsening of sediment lithology that starts in the Pliocene (2–5 Myr ago)²¹. The surface of the 2-km-high plateau consists of Archaean rocks, and equivalent rocks lie 4–5 km below sea level to the north and south of the plateau²² (Fig. 3). The mean surface level of the Shillong plateau is approximately horizontal (Fig. 2c), indicating that the cumulative vertical contributions of slip on the Dauki and Oldham faults have been similar since the plateau began to emerge. If we assume that the relative vertical displacement between the top of the Archaean rocks on the plateau, and their equivalents to north and south, has been caused by slip on those faults over the past 2–5 Myr, then a rate of vertical displacement of 2.5 ± 1 mm yr⁻¹ is implied on each of the faults. Assuming they each dip at 50° , these rates translate into fault slip rates of 3.3 ± 1.3 mm yr⁻¹ and to horizontal shortening rates of 4 ± 2 mm yr⁻¹ on the two faults combined. These estimates are consistent with 1997–99 GPS measurements that indicate the central Shillong plateau moves south at a rate of 6.3 ± 3.8 mm yr⁻¹ relative to points in central and southern India²³.

From these rates we estimate a recurrence interval for earthquakes resembling the 1897 event (slip of 15 m) to be 3–8 kyr on each fault. Palaeoseismic investigations along the northern edge of the plateau indicate a 500-year interval between the past four high-intensity shaking events²⁴, but these may record accelerations from moderate local earthquakes and large events in the Bhutan Himalaya, in addition to events of 1897 type.

Our finding of the reverse slip on the Oldham fault ends a century of speculation as to the mechanism of the 1897 earthquake. The Shillong plateau is not being built as part of a system of thin-skinned thrusting, but is bounded by a high-angle reverse fault to its north, and probably also to its south. The Shillong plateau thus resembles the pop-up structures that border thrust belts elsewhere²⁵. The Shillong faults, and the smaller high-angle compressional faulting in the foreland of the entire Himalayan arc, are located in the region flexed by the load of the Himalaya²⁶. This association implies that the mechanical cause of the deformation is a superposition of elastic stresses caused by bending of the Indian plate and in-plane compressional stress from India's collision with Tibet. As pointed out by Rogers²⁵, most if not all of such structures are, or were when active, near to large mountain chains that could have provided the same state of stress that drives the uplift of the Shillong plateau.

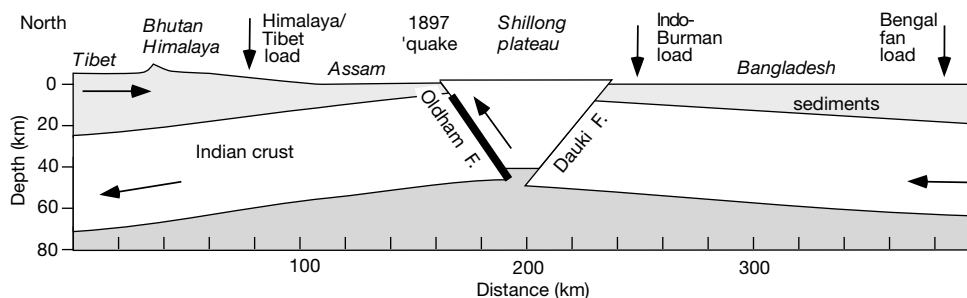


Figure 3 North–south section from Tibet to the Bay of Bengal showing schematic geometry of plateau pop-up. Although the dip of the Dauki fault is conjectural, it would intersect the 1897 rupture were its dip less than $\sim 40^\circ$, resulting in a geometry that is

incompatible with continued slip on both faults. Slip is inhibited³² on reverse faults with dip greater than 50° .

There is thus no need to invoke basal tractions due to subduction to explain pop-up structures, as has often been done for the western Americas^{27,28,29}.

The absence of slip beneath the Bhutan Himalaya in 1897 suggests that the 400-km region between the great Himalayan ruptures of 1934 and 1950 (Fig. 1) has remained a seismic gap for at least the past two centuries³⁰. At the higher end of our estimated slip rates, the faults bounding the Shillong plateau could absorb one-third of the inferred Himalayan contraction rate of 18 mm yr⁻¹ (ref. 31), correspondingly increasing the interval between great earthquakes in the Bhutan Himalaya.

Our conclusions also raise issues concerning the seismic hazard potential of the Shillong plateau. The >300-km length of the Dauki fault has not slipped recently, but were it to slip in a single earthquake its potential maximum magnitude ($M \geq 8$) would constitute a significant seismic threat to nearby densely populated regions of Bangladesh, and to the very large city of Dhaka less than 150 km to the south (Fig. 1). The interval between these giant plateau-building earthquakes fortunately exceeds 3,000 years. □

Received 1 September 2000; accepted 31 January 2001.

- Oldham, R. D. Report on the Great Earthquake of 12 June 1897. *Mem. Geol. Soc. India* **29**, 1–379 (1899).
- Seeber, L. & Armbruster, J. in *Earthquake Prediction: An International Review* (eds Simpson, D. W. & Richards, P. G. 259–277 (Maurice Ewing Series, Vol. 4, American Geophysical Union, Washington DC, 1981).
- Molnar, P. & Pandey, M. R. Rupture zones of great earthquakes in the Himalayan Region. *Proc. Indian Acad. Sci. (Earth Planet. Sci.)* **98**, 61–70 (1989).
- Molnar, P. The distribution of intensity associated with the great 1897 Assam earthquake and bounds on the extent of the rupture zone. *J. Geol. Soc. India* **30**, 13–27 (1987).
- Gahalaut, V. K. & Chander, R. A rupture model for the great earthquake of 1897, northeast India. *Tectonophysics* **204**, 163–174 (1992).
- Bond, J. in *Annual Report of Triangulation 1897–1898* (ed. Strahan, C.) Part II, xii–xiii (Survey of India Department, Calcutta, 1899).
- Burrard, S. G. *Great Trigonometrical Survey of India, North East Longitudinal Series, Synoptical Volume 35* (Survey of India, Dehra Dun, 1909).
- Wilson, C. A. K. *Triangulation of the Assam Valley Series, Geodetic Report 1938*, 10–22 (Survey of India, Dehra Dun, 1939).
- Strahan, G. *Great Trigonometrical Survey of India, Assam Valley Triangulation, Synoptical Volume 32* (Survey of India, Calcutta, 1891).
- Nagar, V. K., Singh, A. N. & Prakesh, A. Strain pattern in N. E. India inferred from geodetic triangulation data. *Mem. Geol. Soc. India* **23**, 265–273 (1992).
- Frank, F. C. Deduction of earth strains from survey data. *Bull. Seismol. Soc. Am.* **56**, 35–42 (1966).
- Okada, Y. Surface deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.* **75**, 1135–1154 (1985).
- Das, J. D., Saraf, A. K. & Jain, A. K. Fault tectonics of the Shillong Plateau and adjoining regions, northeast India, using remote sensing. *Int. J. Remote Sensing* **16**, 1633–1646 (1995).
- Gomberg, J. & Ellis, M. Topography and tectonics of the central New Madrid seismic zone: results of numerical experiments using a three-dimensional boundary element program. *J. Geophys. Res.* **99**, 20299–20310 (1994).
- Chen, W.-P. & Molnar, P. Source parameters of earthquakes and intraplate deformation beneath the Shillong Plateau and the northern Indoburman ranges. *J. Geophys. Res.* **95**, 12527–12552 (1990).
- Le Dain, A. Y., Tappinier, P. & Molnar, P. Active faulting and tectonics of Burma and surrounding regions. *J. Geophys. Res.* **89**, 453–472 (1984).
- Khattari, K. N. Seismological investigations in north eastern region of India. *Mem. Geol. Surv. India* **23**, 275–302 (1992).
- Verma, R. K. & Mukhopadhyay, M. An analysis of the gravity field in northeastern India. *Tectonophysics* **42**, 283–317 (1977).
- Ambraseys, N. Reappraisal of North Indian earthquakes at the turn of the 20th century. *Current Sci.* **79**, 1237–1250 (2000).
- Murthy, M. V. N., Talukdar, S. C., Bhattacharya, A. C. & Chakrabarti, C. The Dauki Fault of Assam. *Bull. Oil Natural Gas Commission* **6**, 57–64 (1969).
- Johnson, S. J. & Alam, A. M. N. Sedimentation and tectonics of the Sylhet trough, Bangladesh. *Geol. Soc. Am. Bull.* **103**, 1513–1527 (1991).
- Evans, P. The tectonic framework of Assam. *J. Geol. Soc. India* **5**, 80–96 (1964).
- Paul, J. et al. Active deformation across India. *Geophys. Res. Lett.* **28**, 647–651 (2001).
- Sukhija, B. S. et al. Timing and return of major paleoseismic events in the Shillong Plateau, India. *Tectonophysics* **308**, 53–65 (1999).
- Rogers, J. Chains of basement uplifts within cratons marginal to orogenic belts. *Am. J. Sci.* **287**, 661–692 (1987).
- Lyon-Caen, H. & Molnar, P. Gravity anomalies, flexure of the Indian Plate, and the structure, support and evolution of the Himalaya and Ganga basin. *Tectonics* **4**, 513–538 (1985).
- Lipman, P. W., Prottska, H. J. & Christiansen, R. L. Evolving subduction zones in the western United States, as interpreted from igneous rocks. *Science* **174**, 821–825 (1971).
- Dickinson, W. R. & Snyder, W. S. Plate tectonics of the Laramide orogeny. *Geol. Soc. Am. Mem.* **151**, 355–366 (1978).
- Jordan, T. E. et al. Andean tectonics related to the geometry of the subducted Nazca plate. *Geol. Soc. Am. Bull.* **94**, 341–361 (1983).
- Bilham, R. & Gaur, V. K. The geodetic contribution to Indian seismotectonics. *Current Sci.* **79**, 1259–1269 (2000).

- Bilham, R., Larson, K., Freymueller, J., & Project Idylhim members. GPS measurements of present-day convergence across the Nepal Himalaya. *Nature* **386**, 61–64 (1997).
- Sibson, R. H. & Xie, G. Dip range for intracontinental reverse fault ruptures: the truth not stranger than friction. *Bull. Seismol. Soc. Am.* **88**, 1014–1022 (1998).

Acknowledgements

This work was funded by the National Science Foundation and the Natural Environment Research Council. R.B. received a John Simon Guggenheim Memorial Foundation fellowship while at Oxford University.

Correspondence and requests for materials should be addressed to R.B. (e-mail: bilham@stripe.colorado.edu).

Plant diversity enhances ecosystem responses to elevated CO₂ and nitrogen deposition

Peter B. Reich*, Jean Knops†, David Tilman‡, Joseph Craine‡, David Ellsworth§, Mark Tjoelker*, Tali Lee*, David Wedin||, Shahid Naeem†, Dan Bahaiddin*, George Hendrey§, Shibu Jose*, Keith Wrage*, Jenny Goth* & Wendy Bengtson*

* Department of Forest Resources, University of Minnesota, St Paul, Minnesota 55108, USA

† Department of Ecology, Evolution and Behavior, University of Minnesota, St Paul, Minnesota 55108, USA

‡ Department of Integrative Biology, University of California, Berkeley, California 94720, USA

§ Division of Environmental Biology, Brookhaven National Laboratory, Upton, New York 11973, USA

|| School of Natural Resource Sciences, University of Nebraska, Lincoln, Nebraska 68583, USA

Human actions are causing declines in plant biodiversity, increases in atmospheric CO₂ concentrations and increases in nitrogen deposition; however, the interactive effects of these factors on ecosystem processes are unknown^{1,2}. Reduced biodiversity has raised numerous concerns, including the possibility that ecosystem functioning may be affected negatively^{1–4}, which might be particularly important in the face of other global changes^{5,6}. Here we present results of a grassland field experiment in Minnesota, USA, that tests the hypothesis that plant diversity and composition influence the enhancement of biomass and carbon acquisition in ecosystems subjected to elevated atmospheric CO₂ concentrations and nitrogen deposition. The study experimentally controlled plant diversity (1, 4, 9 or 16 species), soil nitrogen (unamended versus deposition of 4 g of nitrogen per m² per yr) and atmospheric CO₂ concentrations using free-air CO₂ enrichment (ambient, 368 μmol mol⁻¹, versus elevated, 560 μmol mol⁻¹). We found that the enhanced biomass accumulation in response to elevated levels of CO₂ or nitrogen, or their combination, is less in species-poor than in species-rich assemblages.

In the twenty-first century humans will live in, manage and depend on ecosystems that are less diverse^{1,2} and subjected to higher CO₂ levels and nitrogen (N) deposition rates than in recorded human history¹. Although we are beginning to understand the individual impacts of each of these factors on terrestrial ecosystems, our understanding of their interactive effects is poor at best^{1–10}. Net primary productivity and carbon (C) input to ecosystems are usually enhanced by elevated CO₂, but this seems to be related to the extent and type of other limitations^{11–20}. Nitrogen-poor ecosystems have often shown less response to elevated CO₂ than more fertile systems^{18–25}, which is important given that, worldwide,

productivity in most terrestrial ecosystems is limited by N (ref. 26) and rates of N deposition are expected to increase in the future^{1,2,6}. In addition, it has been proposed that species-poor communities may be less responsive to elevated CO₂ or N compared with diverse communities⁷.

Our study focuses on the influence of plant diversity on the response of ecosystem functions, such as productivity, to elevated CO₂ and N supply rates. Theoretically, because the range in plant traits associated with acquisition of C and N is frequently correlated with the species or functional diversity of an ecosystem^{7,10}, ecosystem responses to elevated CO₂ and N deposition may be sensitive to variation in levels of biodiversity^{4,7,10,11,13–19}. High-diversity plant communities frequently have a greater range of plant functional traits that affect C (for example, C₃ and C₄ photosynthetic pathways) and N (for example, legumes associated with N-fixing symbionts and non-leguminous species) cycling than have species-poor communities^{7,10}, and therefore have potentially greater responsiveness of ecosystem functions to elevated CO₂ and N deposition. This can occur because diverse assemblages have a greater likelihood of containing species with strong responses to resources and strong impacts on ecosystem processes compared with species-poor assemblages (that is, a sampling effect)^{27,28}. Alternatively, this can occur because the greater range of traits extant in diverse assemblages positively affects competitive associations and interactions (such as niche complementarity and positive species interactions)^{3,7,10,28}.

Testing the influence of plant diversity on the impacts of elevated CO₂ and N to ecosystem functioning requires simultaneous manipulation of plant species composition, CO₂, and N. Although there have been glasshouse studies⁸ of species responses to elevated CO₂ in monocultures versus mixtures, and field studies of monocultures and bi-species mixtures^{9,15}, this is the first field study, to our knowledge, to test the hypothesis⁷ that plant species diversity influences ecosystem-scale biomass responses to elevated CO₂ and N levels. The BioCON experimental facility (see Methods) in Minnesota, USA, was designed expressly for the simultaneous manipulation of these three factors in experimental grassland plots under field conditions, using a well-replicated split-plot experiment comprising a full factorial combination of treatment levels in a completely randomized design.

Our study includes 296 individual plots (each 2 × 2 m) distributed among six 20-m diameter experimental areas (rings). In three elevated CO₂ rings, a free-air CO₂ enrichment system²⁹ was used during the 1998 and 1999 growing seasons to maintain the CO₂ concentration at an average of 560 μmol mol⁻¹, a concentration likely to be reached this century^{1,2,6}. Three ambient rings (368 μmol mol⁻¹ CO₂) were treated identically but without additional CO₂. All plots were planted with 1, 4, 9 or 16 perennial grassland plant species randomly chosen from 16 species in 4 functional groups (C₃ grasses, C₄ grasses, C₃ legumes, C₃ non-legume forbs). Hence high-diversity treatments incorporate greater species and functional group diversity. Native and secondary grasslands in Minnesota typically contain a mixture of these four functional types. Beginning in 1998, half the plots received addi-

tional N equivalent to 4 g N m⁻² yr⁻¹, which is comparable to high rates of N deposition observed in industrialized regions¹.

Diversity, CO₂ and N treatments had significant main effects on total biomass (Table 1, Fig. 1). Above-ground and below-ground biomass both increased markedly with increasing species diversity. Elevated CO₂ predominantly stimulated below-ground biomass, whereas enriched N largely increased above-ground biomass (Table 1). Elevated CO₂ decreased and enriched N increased plant N concentration, total plant N, soil net N mineralization and soil solution N pools (Table 1).

There was no significant interaction between CO₂ and N, and elevated levels of both resources increased total biomass by 27% on average compared with ambient plots (Fig. 1). In essence, the productivity response to elevated CO₂ was not constrained by N limitation in unfertilized plots, even those lacking N-fixing legumes (data not shown). This result differs from many earlier studies^{11,18,20}, perhaps because we added smaller amounts of N (4 g N m⁻² yr⁻¹) to represent elevated N deposition, whereas other studies (such as ref. 18) added high amounts (up to 56 g N m⁻² yr⁻¹) to mimic agricultural N addition.

Although we did not detect a CO₂ and N interaction, there were significant biomass interactions between diversity and CO₂ as well as between diversity and N (Fig. 1; and Methods), largely owing to responses of below-ground biomass, which comprised three-quarters of the total biomass. As the response to elevated CO₂ and N varied among diversity treatments, we tested their effects within diversity levels, using post-hoc comparisons of the mean responses averaged over all four harvests. In the 16-species plots, all treatments

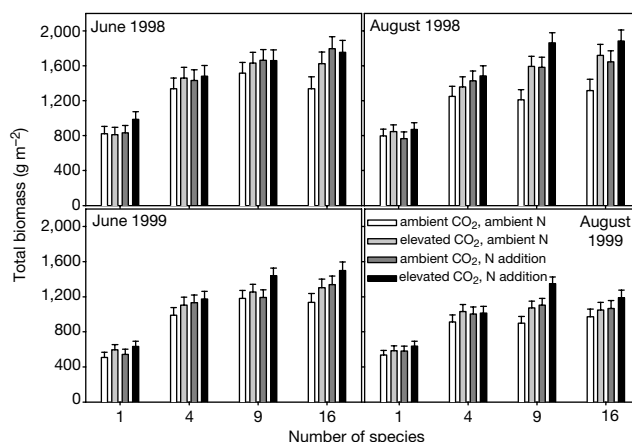


Figure 1 Total biomass (above-ground plus below-ground, 0–20 cm depth) (± 1 s.e.) for plots planted with either 1, 4, 9 or 16 species, grown at four combinations of ambient (368 $\mu\text{mol mol}^{-1}$) and elevated (560 $\mu\text{mol mol}^{-1}$) concentrations of CO₂, and ambient N and N addition (4 g N m⁻² yr⁻¹). Biomass data are shown for each of four harvests (June and August in both 1998 and 1999). The biomass (total and/or below-ground) response to elevated CO₂, enriched N, or both, differed significantly among diversity treatments at every harvest.

Table 1 Average two-year responses of experimental grassland communities to elevated atmospheric CO₂ concentrations, N deposition, and species diversity

Parameter	R ²	F ratio	CO ₂ % change	N % change	Species diversity % change
Total biomass	0.54	16.9	+12**	+13***	+98***
Above-ground biomass	0.39	9.2	+8	+23***	+81***
Below-ground biomass	0.50	15.0	+14*	+9*	+105***
Whole plant %N	0.29	5.9	-13***	+14***	-26***
Total plant N content	0.47	12.9	-2	+29***	+64***
Soil net N mineralization	0.23	4.2	-15	+64*	-68***
Soil solution N (0–20 cm)	0.21	3.9	-14	+36	-76**

R² and F ratio shown for the whole model (see Methods), for which P values were always less than 0.001. Main effects (% difference, pooled across all other treatments) and significant level ($\dagger$, $P < 0.1$; *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$) shown are based on measurements in all 296 plots. Species diversity % effects shown compare 16-species versus 1-species plots.

with elevated levels of either CO₂ or N had significantly greater biomass than the ambient CO₂/ambient N plots. In the 9-species plots, only the elevated CO₂/elevated N treatment had significantly greater biomass than the ambient/ambient treatment. In contrast, there was no significant effect of CO₂ or N treatment, alone or in combination, on total biomass for either the 1- or 4-species levels.

Thus, across the four harvests in 1998 and 1999, the enhancement of biomass owing to either elevated CO₂ or enriched N decreased with declining diversity. In the unamended N treatment, the average stimulation of total biomass in response to elevated CO₂ was 22% in 16-species plots, 18% in 9-species plots, 10% in 4-species plots and 7% in monoculture plots (Fig. 2). Differences in plant species diversity accounted for a fivefold difference in the impacts of CO₂ fertilization on biomass accumulation. For example, whereas elevated CO₂ increased biomass by 258 g m⁻² on average in the most diverse (16-species) plots, it increased biomass by only 47 g m⁻² on average in the monocultures (Fig. 2).

Similarly, under ambient CO₂, the average enhancement of total biomass in response to N addition ranged from 25% in 16-species plots to 18%, 11% and 2% in 9-species, 4-species and monoculture plots, respectively (Fig. 2). Thus, biomass enhancement by N enrichment was much less in low-diversity plots—N deposition increased biomass by almost 300 g m⁻² in the diverse plots and by only 15 g m⁻² in the monocultures. For plots subjected to both elevated CO₂ and enriched N, biomass increased by more than 400 g m⁻² in both the 16- and 9-species plots (+35%), but only by

roughly 100–150 g m⁻² in 4-species plots (+15%) and monocultures (+17%).

By sorting above-ground biomass to species, we found that the increase in biomass with increased diversity at each resource level was due largely to the four species that were most abundant in mixtures. The four species *Achillea*, *Bromus*, *Lupinus* and *Poa* together made up more than 80% of the above-ground biomass in 16-species plots (Table 2) and represent all three of the C₃ functional groups. Three of these species were among the four most productive species in monocultures. However, all four of the species abundant in mixtures had lower average monoculture biomass at a given resource level than the average 16-species mixtures at the same resource level (Table 2) and no one species comprised more than one-third of the above-ground biomass in the 16-species mixture at any level. Thus, consistent with a multiple-species 'sampling effect', there was collective dominance in 16-species mixtures by three of the more productive members of the species pool, but this dominance was moderate, and other species also contributed to the greater biomass of the high-diversity plots.

The effects of species diversity on biomass accumulation responses to elevated CO₂ and N deposition also appear to result from the combination of multiple-species sampling effects and niche complementarity or positive species interactions. The four dominant species (*Achillea*, *Bromus*, *Lupinus* and *Poa*) were responsible for much of the increase in biomass in 16-species plots at elevated resource levels (that is, elevated CO₂, enriched N, or both) and showed large responses to these treatments when grown in monoculture (Table 2; for example, *Achillea* and *Lupinus* for elevated CO₂; *Poa* and *Bromus* for elevated N).

Although the response of *Poa* monocultures to N fertilization was large enough to suggest that it may have the capability of explaining most of the increased biomass response of the 16-species mixture if heavily dominant, *Poa* was not more than one-third of the total above-ground biomass in those mixtures, indicating that other species are also important. For elevated CO₂ at ambient N, the stimulation of total biomass (+258 g m⁻²) in 16-species mixtures was greater than the stimulation in monoculture of any species to elevated CO₂. Thus, the increase in biomass of 16-species plots with increasing resource supply was substantially due to the presence and response of species that do respond strongly in monocultures (that is, interpretable as mainly a sampling effect), but they collectively increased more when together than when in monoculture (a niche effect).

In addition, different species were responsible for the enhanced responses to different combinations of CO₂ and N, again suggesting niche differentiation. The increased total biomass response of 16-species plots to elevated CO₂ at ambient N (of +22%, Figs 1 and 2) was largely due to the response of *Achillea*, *Lupinus* and *Bromus*, each of which maintained a roughly similar fraction of a higher total above-ground biomass under elevated than ambient CO₂ (Table 2). At either ambient or elevated CO₂, the enhanced biomass in N-enriched 16-species plots as compared with ambient/ambient plots was due to increases in biomass by *Poa*, *Bromus* and *Achillea*, in that

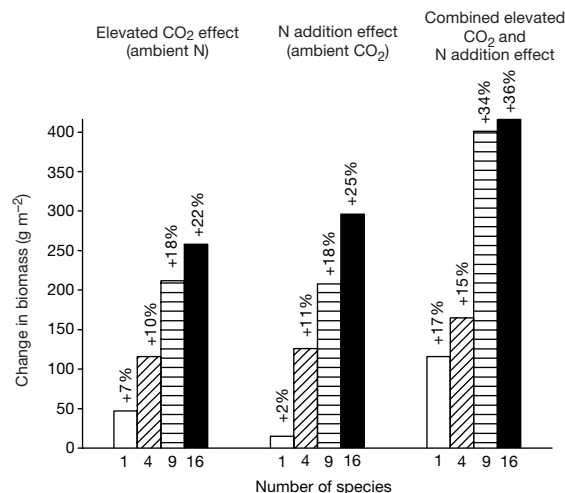


Figure 2 Change in total (above-ground plus 0–20 cm below-ground) biomass (compared with ambient levels of both CO₂ and N) in response to elevated CO₂ alone (at ambient soil N), to enriched N alone (at ambient CO₂), and to the combination of elevated CO₂ and enriched soil N, for plots containing 1, 4, 9 or 16 species. Data were averaged for 4 harvests over 2 yr. Per cent change is shown above each histogram for each diversity treatment.

Table 2 Total biomass of 16-species mixtures and of monocultures of the dominant species in the 16-species mixtures

	Ambient CO ₂ , ambient N		+CO ₂ , ambient N		Ambient CO ₂ , +N		+CO ₂ , +N	
Biomass (g m ⁻²) of 16-species mixtures	1,165		1,423		1,462		1,581	
Species	Monoculture biomass (g m ⁻²)	% of 16-species mixture	Monoculture biomass (g m ⁻²)	% of 16-species mixture	Monoculture biomass (g m ⁻²)	% of 16-species mixture	Monoculture biomass (g m ⁻²)	% of 16-species mixture
<i>Achillea</i>	1,124	34%	1,314	33%	848	31%	1,321	34%
<i>Poa</i>	979	22%	1,089	19%	1,322	33%	1,407	27%
<i>Lupinus</i>	509	15%	672	17%	449	6%	670	9%
<i>Bromus</i>	1,021	11%	989	12%	1,228	14%	1,082	14%
All other species	590	18%	615	19%	587	16%	664	16%

Totals are above-ground plus below-ground biomass, at all four combinations of CO₂ and N. Also shown is the % of total above-ground biomass for each of the dominant species in the 16-species mixtures at each CO₂ and N combination.

order, despite a decline in *Lupinus* biomass.

These results suggest that multiple-species sampling effects, niche complementarity and positive species interactions^{3,7,10,28} jointly help to explain the greater responses of diverse than species-poor plots to elevated levels of the two major global change factors: atmospheric CO₂ and N deposition. Moreover, the central finding of this study—that changes in plant diversity influence the magnitude of CO₂ and N impacts on ecosystem functioning—is important regardless of which set of mechanisms are operating.

Our study raises concerns about the consequences of widespread changes in plant composition and diversity in ecosystems worldwide for responses to other global changes^{1,2}, because the results show that, in response to elevated levels of CO₂ and N, ecosystems with decreased diversity may acquire less C and biomass than ecosystems with greater diversity. It is unclear whether the responses observed in 1998–1999 will change with time as the composition of the communities changes, the climate varies, and long-term soil microbial feedbacks occur^{20–25}, or whether natural or managed vegetation would respond similarly to combinations of diversity, CO₂ and N. Nonetheless, our results suggest that the reduction of diversity occurring globally may reduce the capacity of ecosystems to capture additional C under conditions of rising atmospheric CO₂ concentrations and N deposition levels. □

Methods

BioCON experiment

The BioCON (Biodiversity, CO₂ and N) experiment (<http://swan.lter.umn.edu/biocon/>) is located at the Cedar Creek Natural History area in Minnesota, USA. Plots were established on a secondary successional grassland on a sandy outwash soil after removing the previous vegetation. The experimental treatments were arranged in complete factorial combination of CO₂ (ambient or 560 µmol mol⁻¹), species number (1, 4, 9, and 16) and N level (control and fertilized). The species numbers were chosen as the squares of 1, 2, 3 and 4 to represent roughly equal effective differences in diversity, on the basis of earlier studies⁴. Each plot was planted in 1997 with 12 g m⁻² of seed partitioned equally among all species planted in a plot.

The design consisted of a split-plot arrangement of treatments in a completely randomized design. CO₂ treatment is the whole-plot factor and is replicated three times among the six rings. The subplot factors of species number and N treatment were assigned randomly and replicated in individual plots among the six rings. For each of the four combinations of CO₂ and N levels, pooled across all rings, there were 32 randomly assigned replicates for the plots planted to 1 species, 15 for those planted to 4 species, 15 for 9 species, and 12 for 16 species. Beginning in 1998, the plots assigned to the N addition treatment were amended with 4 g N m⁻² yr⁻¹, applied over three dates each year. CO₂ was added in elevated treatments during all daylight hours from 9 April to 16 October 1998, and from 20 April to 9 November 1999. Although there was modest variation in CO₂ concentrations spatially within and across rings, the average CO₂ concentrations were not more than 1–2 µmol mol⁻¹ different among rings, or among diversity or N treatment levels averaged within and across rings (see Supplementary Information).

Species and biomass measurements

The 16 species used in this study were all native or naturalized to the Cedar Creek Natural History Area. They include four C₄ grasses (*Andropogon gerardii*, *Bouteloua gracilis*, *Schizachyrium scoparium*, *Sorghastrum nutans*), four C₃ grasses (*Agropyron repens*, *Bromus inermis*, *Koeleria cristata*, *Poa pratensis*), four N-fixing legumes (*Amorpha canescens*, *Lespedeza capitata*, *Lupinus perennis*, *Petalostemum villosum*) and four non-N-fixing herbaceous species (*Achillea millefolium*, *Anemone cylindrica*, *Asclepias tuberosa*, *Solidago rigida*), and all are referred to by genus elsewhere. Monocultures of all species were replicated twice at all CO₂ and N levels. The 4- and 9-species plots were random selections from all species. Plots were regularly weeded to remove unwanted species. In June and August of each year, we assessed above- and below-ground (0–20 cm) biomass, plant C and N, and soil N (see Supplementary Information). Soil net N mineralization rates were measured once each year. Above-ground biomass was sorted to species at each harvest. The species richness of clipped above-ground biomass samples was 1.0, 3.8, 8.0 and 13.6 species on average for plots planted with 1, 4, 9 and 16 species, respectively, on the basis of the number of species identified in each plot from the four above-ground harvests (each 0.1 m²).

Statistical analysis

In analysis of variance all treatment effects were considered fixed. Using *F*-tests, the effect of CO₂ (1 degree of freedom, d.f.) was tested against the random effect of ring nested within CO₂ (4 d.f.). The main effects of species number (3 d.f.) and N (1 d.f.), and interactions between CO₂ and N were tested against the residual error. The main effect of species number and its first-order interaction terms were partitioned into single-degree-of-freedom contrasts for linear, quadratic and cubic terms to test for interactions between diversity and either CO₂ or N treatments, and additionally to test hypotheses about

predetermined contrasts of elevated resource levels (elevated CO₂, enriched N, or both) versus the ambient/ambient conditions. There were interactions for below-ground biomass between species diversity and CO₂ treatments in the August 1998 (*P* < 0.05), June 1999 (*P* < 0.10) and August 1999 (*P* < 0.10) harvests, and between species diversity and N treatments in the June 1999 (*P* < 0.05) and August 1999 (*P* < 0.05) harvests. For the pre-planned contrast between the ambient/ambient and the elevated CO₂/enriched N treatments, there were significantly different responses (for below-ground and total biomass) for different diversity treatments at the June 1998 (*P* < 0.05), August 1998 (*P* < 0.05) and June 1999 (*P* < 0.05) harvests. Given interactions between species diversity and resource treatments, post-hoc Student's *t*-tests were also conducted to compare individual resource treatments within species diversity levels. All analyses were conducted for each harvest, for each year (pooling harvest data by plot), and across years averaged (pooling data for all harvests by plot), with similar results, although there was harvest-to-harvest variation (see Fig. 1).

Received 20 November 2000; accepted 18 January 2001.

1. Vitousek, P. M. Beyond global warming: ecology and global change. *Ecology* **75**, 1861–1876 (1994).
2. Sala, O. E. *et al.* Global biodiversity scenarios for the year 2100. *Science* **287**, 1770 (2000).
3. Hector, A. *et al.* Plant diversity and productivity experiments in European grasslands. *Science* **286**, 1123 (1999).
4. Tilman, D. *et al.* The influence of functional diversity and composition on ecosystem processes. *Science* **277**, 1300–1302 (1997).
5. Stocker, R., Körner, C., Schmid, B., Niklaus, P. A. & Leadley, P. W. A field study of the effects of elevated CO₂ and plant species diversity on ecosystem-level gas exchange in a planted calcareous grassland. *Global Change Biol.* **5**, 95–105 (1999).
6. Schimel, D. *et al.* Contribution of increasing CO₂ and climate to carbon storage by ecosystems in the United States. *Science* **287**, 2004 (2000).
7. Bolker, B. M., Pacala, S. W., Bazzaz, F. A., Canham, C. D. & Levin, S. A. Species diversity and ecosystem response to carbon dioxide fertilization: conclusions from a temperate forest model. *Global Change Biol.* **1**, 373 (1995).
8. Navas, M.-L., Garnier, E., Austin, M. P. & Gifford, R. M. Effect of competition on the responses of grasses and legumes to elevated atmospheric CO₂ along a nitrogen gradient: differences between isolated plants, monocultures and multi-species mixtures. *New Phytol.* **143**, 323–331 (1999).
9. Schenk, U., Jäger, H.-J. & Weigel, H.-J. The response of perennial ryegrass/white clover swards to elevated atmospheric CO₂ concentrations. *New Phytol.* **135**, 67–79 (1997).
10. Schmid, B., Joshi, J., Schläpfer, F. in *Functional Consequences of Biodiversity: Experimental Progress and Theoretical Extensions* (eds Kinzig, A., Tilman, D. & Pacala, S.) (Princeton Univ. Press, Princeton, in the press).
11. Körner, C. & Bazzaz, F. A. *Carbon Dioxide, Populations and Communities* (Academic, San Diego, 1996).
12. DeLucia, E. H. *et al.* Net primary production of a forest ecosystem with experimental CO₂ enrichment. *Science* **284**, 1177–1179 (1998).
13. Wand, S. J., Midgley, G. F., Jones, M. H. & Curtis, P. S. Responses of wild C₄ and C₃ (Poaceae) species to elevated atmospheric CO₂ concentration: a meta-analytic test of current theories and perceptions. *Global Change Biol.* **5**, 723–741 (1999).
14. Owensby, C. E., Ham, J. M., Knapp, A. K. & Auen, L. M. Biomass production and species composition change in a tallgrass prairie ecosystem after long-term exposure to elevated atmospheric CO₂. *Global Change Biol.* **5**, 497–506 (1999).
15. Lüscher, A., Hendrey, G. R. & Nösberger, J. Long-term responsiveness to free-air CO₂ enrichment of functional types, species, and genotypes of plants from fertile permanent grassland. *Oecologia* **113**, 37–45 (1998).
16. Warwick, K. R., Taylor, G. & Blum, H. Biomass and compositional changes occur in chalk grassland turves exposed to elevated CO₂ for two seasons in FACE. *Global Change Biol.* **4**, 375–385 (1998).
17. Leadley, P. W., Niklaus, P. A., Stocker, R. & Körner, C. A field study of the effects of elevated CO₂ on plant biomass and community structure in a calcareous grassland. *Oecologia* **118**, 39–49 (1999).
18. Zanetti, S. *et al.* Does nitrogen nutrition restrict the CO₂ response of fertile grassland lacking legumes? *Oecologia* **112**, 17–25 (1997).
19. Hebeisen, T. *et al.* Growth response of *Trifolium repens* L. and *Lolium perenne* L. as monocultures and bi-species mixture to free air CO₂ enrichment and management. *Global Change Biol.* **3**, 149–160 (1997).
20. Zak, D. R. *et al.* Atmospheric CO₂, soil-N availability, and allocation of biomass and nitrogen by *Populus tremuloides*. *Ecol. Appl.* **10**, 34–46 (2000).
21. Hu, S., Firestone, M. K., Chapin, F. S. III. Soil microbial feedbacks to atmospheric CO₂ enrichment. *Trends Ecol. Evol.* **14**, 433 (1999).
22. Diaz, S., Grime, J. P., Harris, J. & McPherson, E. Evidence of a feedback mechanism limiting plant response to elevated carbon dioxide. *Nature* **364**, 616–617 (1993).
23. Zak, D. R. *et al.* Elevated atmospheric CO₂ and feedback between carbon and nitrogen cycles. *Plant Soil* **151**, 105–117 (1993).
24. Hungate, B. A. *et al.* The fate of carbon in grasslands under carbon dioxide enrichment. *Nature* **388**, 576–579 (1997).
25. Cannell, M. & Thornley, J. N-poor ecosystems may respond more to elevated [CO₂] than N-rich ones in the long term. A model analysis of grassland. *Global Change Biol.* **4**, 431–442 (1998).
26. Vitousek, P. M. & Howarth, R. W. Nitrogen limitation on land and in the sea: how can it occur? *Biogeochemistry* **13**, 87–115 (1991).
27. Huston, M. A. Hidden treatments in ecological experiments: re-evaluating the ecosystem function of biodiversity. *Oecologia* **110**, 449–460 (1997).
28. Tilman, D., Lehman, C. L. & Thomson, K. T. Plant diversity and ecosystem productivity: theoretical considerations. *Proc. Natl Acad. Sci. USA* **94**, 1857–1861 (1997).
29. Lewin, K. F., Hendrey, G. R., Nagy, J. & LaMorte, R. Design and application of a free-air carbon dioxide enrichment facility. *Agric. Forest Meteorol.* **70**, 15–29 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Correspondence and requests for materials should be addressed to P.R. (e-mail: preich@forestry.umn.edu).

Chemical speciation drives hydrothermal vent ecology

George W. Luther III*, Tim F. Rozan*, Martial Tallefer†, Donald B. Nuzzio‡, Carol Di Meo*, Timothy M. Shank§, Richard A. Lutz§ & S. Craig Cary*

* College of Marine Studies, University of Delaware, Lewes, Delaware 19958, USA

† Analytical Instrument Systems, Inc., 1059C Old York Road, Ringoes, New Jersey 08851, USA

‡ Biology Department, MS #34 1-16 Redfield, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

§ Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, New Jersey 08901, USA

The physiology and biochemistry of many taxa inhabiting deep-sea hydrothermal vents have been elucidated¹⁻⁴; however, the physicochemical factors controlling the distribution of these organisms at a given vent site remain an enigma after 20 years of research⁵⁻¹¹. The chemical speciation of particular elements has been suggested as key to controlling biological community structure in these extreme aquatic environments^{7,11,12}. Implementation of electrochemical technology^{13,14} has allowed us to make *in situ* measurements of chemical speciation at vents located at the East Pacific Rise (9° 50' N) and on a scale relevant to the biology. Here we report that significant differences in oxygen, iron and sulphur speciation strongly correlate with the distribution of specific taxa in different microhabitats. In higher temperature (> 30 °C) microhabitats, the appreciable formation of soluble iron-sulphide molecular clusters markedly reduces the availability of free H₂S/HS⁻ to vent (micro)organisms, thus controlling the available habitat.

To determine the availability of biologically relevant chemical compounds in mixing zones from diffuse flow venting areas, *in situ* chemical analyses are essential^{7,11,12}. Earlier *in situ* measurements at hydrothermal vents used flow injection analysis to measure acid-volatile sulphide (AVS = Σ[free H₂S/HS⁻ plus FeS])^{11,15}, but this technique does not permit identification of individual sulphur components. At a given vent location, AVS generally increases with temperature, as expected from conservative mixing; however, no clear chemical pattern that can explain the different environmental niches occupied by different vent organisms has been discerned¹¹. As iron- and sulphide-rich vent fluids meet ambient seawater, dissolved iron reacts with sulphide to form solid FeS ('black smoke')¹⁶, which reduces the availability of free H₂S to organisms. Studies¹⁶⁻¹⁸ have shown that dissolved molecular clusters of FeS (FeS (aq)) can be discriminated from free H₂S/HS⁻ and polysulphides by voltammetry. Here we have used an *in situ* voltammetric sensor package for real-time measurements at deep-sea hydrothermal vents.

Surveys of ambient deep-sea water and vent end-member fluids showed the expected chemistry. Linear sweep voltammograms (LSVs) from ambient seawater revealed only the presence of O₂ (Fig. 1a). Conversely, cyclic voltammograms, taken with the electrodes placed 0.5 m (25 °C) above a black smoker, showed only the presence of free H₂S/HS⁻ and FeS (aq) (Fig. 1b). These results provided the framework for comparative studies of the habitats for different organisms: the tubeworm *Riftia pachyptila* and the polychaete *Alvinella pompejana*. The tubeworm resides in lower temperature fluids, issuing through seafloor openings or from the base of chimneys (< 25 °C), and harbours bacterial endosymbionts^{1,2}, which produce organic carbon for the host by the bacterially

mediated chemosynthetic reaction of H₂S/HS⁻ with CO₂. In contrast, *A. pompejana*, which forms tube dwelling colonies on the sides of active sulphide chimneys (> 25 °C), does not rely on chemosynthetic endosymbionts but directly ingests bacteria in, on and outside its tubes¹⁹.

In a community of living *R. pachyptila*, electrodes were placed near their plumes, where acquisition of H₂S/HS⁻, CO₂, and O₂ occurs for chemoautotrophy. Only H₂S/HS⁻ and O₂ were measured

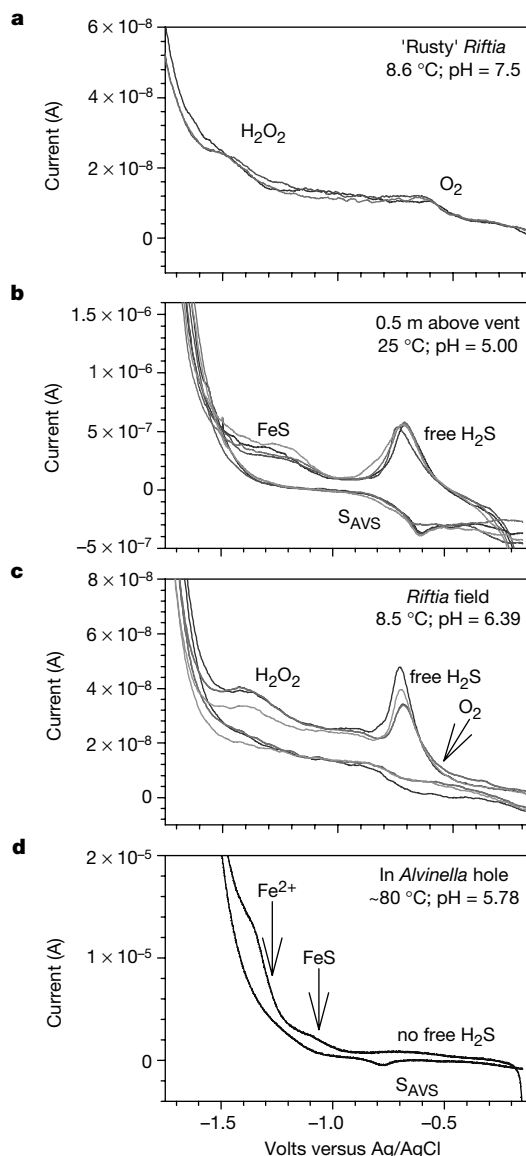


Figure 1 Representative *in situ* voltammetry data. **a**, LSV scans at dead *Riftia* sites or where vent biota do not exist. O₂ reacts to form H₂O₂ at a half-wave potential ($E_{1/2}$) of -0.5 V, and the resulting H₂O₂ reduces to water at -1.30 V. H₂O₂ overlaps with the Fe²⁺ signal so traces of Fe²⁺ cannot be discerned. **b**, Cyclic voltammogram scans taken 0.5 m above a vent orifice showing free H₂S/HS⁻ and FeS (aq). Free H₂S/HS⁻ reacts to form a Hg film at positive potentials and is reduced during the negative scan to release sulphide at $E_{1/2} = -0.65$ V. FeS (aq) gives an irreversible signal ($E_{1/2} = -1.12$ V) owing to reduction of Fe²⁺ in FeS (aq) to Fe(Hg) which decomposes FeS (aq) (ref. 17). On scan reversal, the negative current signal ($E_{1/2} = -0.62$ V) is from all sulphide (AVS = free H₂S/HS⁻ and sulphide released from FeS (aq)) reacting with Hg to reform HgS. **c**, Cyclic voltammogram scans from *Riftia* sites showing only O₂ and free H₂S/HS⁻; the AVS signal is small when only H₂S/HS⁻ is present. **d**, Cyclic voltammogram scan (200 mV s⁻¹) taken at 80 ± 20 °C from within an *Alvinella* tube shows Fe²⁺ ($E_{1/2} = -1.30$ V) and FeS (aq) ($E_{1/2} = -1.05$ V) as the main chemical species. The negative current signal (-0.76 V) results only from the breakdown of FeS (aq). The current axis is much higher because it is a function of temperature.

|| Present address: School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia 30322, USA.

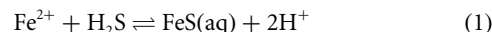
by voltammetry (Fig. 1c), consistent with previous observations¹¹. Another survey from an assemblage of dead *R. pachyptila* showed that surrounding waters contained only O₂ (Fig. 1a), even though the waters (8.5 °C) were above ambient deep-sea temperatures (2 °C). The tubes were also encrusted with orange-brown solids containing Fe(III). Absence of free H₂S/HS⁻, caused by a change in chemistry of emanating vent fluids, indicates that sulphide chemosynthesis cannot occur and support a thriving *R. pachyptila* community.

Chemical speciation of *A. pompejana* habitats was significantly different from live and dead *R. pachyptila* areas. The electrodes, placed directly inside occupied *A. pompejana* tubes, revealed large signals (Fig. 1d) due to FeS (aq) and Fe²⁺. Free H₂S/HS⁻ was not detected, or only at trace levels, with no detectable O₂. Scans taken at tube openings and on the sulphide structure—the substrate for the *A. pompejana* colonies—showed similar iron and sulphur chemistry. These data indicate that chemical speciation is unique to the sulphide chimney structure and is not being modified by the *A. pompejana* colony. Free H₂S/HS⁻ was detected higher above the colony where the temperature approaches that of ambient seawater, suggesting that FeS dissociation occurred.

In addition, discrete samples²⁰ taken from 8 *R. pachyptila* and 11 *A. pompejana* locations (Fig. 2) were analysed aboard ship for AVS, pH and total iron (Fe_T) concentrations. AVS concentrations were at least five times higher at *A. pompejana* sites than at *R. pachyptila* sites, in accord with previous work^{6,10}. H₂S/HS⁻ is the dominant sulphide phase in live *R. pachyptila* fields, and in six of the eight samples, free H₂S/HS⁻ ranged from 68 to 100% of the AVS. In contrast, FeS (aq) is the chief sulphide form in *A. pompejana* tubes and near their tube openings. Free H₂S/HS⁻ ranged from 0 to 39% of the total AVS in *A. pompejana* habitats, and its concentration is typically lower than that measured in *R. pachyptila* fields even though the total AVS is significantly higher.

At higher iron and AVS concentrations where *A. pompejana* resides (Fig. 3), formation of solid FeS and FeS (aq) is enhanced because of higher temperatures^{16,21}. This reduces the free H₂S

concentration while lowering pH to about 5 (eq. 1). This reaction is supported by the significant correlation of Fe_T versus AVS for *A. pompejana* samples ($r^2 = 0.574$; slope Fe_T/AVS = 1.52). There are also significant inverse correlations between Fe_T or AVS versus pH for the *A. pompejana* samples from the different sites (Fig. 3). We propose that FeS (aq) molecular cluster formation is the primary mechanism controlling free sulphide detoxification for *A. pompejana* in vent waters.



In contrast, *R. pachyptila* resides in lower temperature (< 25 °C) and higher pH (> 5.7) waters where there is no correlation of AVS versus pH (Fig. 3) and of Fe_T versus AVS ($r^2 = 0.019$). Thus, FeS formation is insignificant. In *R. pachyptila* fields, we did not observe a signal for FeS (aq), and free H₂S/HS⁻ is the dominant sulphur species. We detected trace amounts of Fe_T in three of the eight locations, and calculations indicated that the K_{sp} (solubility product constant) for FeS was not exceeded in *R. pachyptila* communities.

Unlike *A. pompejana*, *R. pachyptila* requires free sulphide so the bacteria endosymbionts can convert CO₂ to organic matter for translocation to their host. Research indicates that *R. pachyptila* transports and uses HS⁻ preferentially to H₂S (ref. 4). As the first acid dissociation constant (pK_a) of H₂S in seawater is 6.7 (ref. 22), at pH 5.7 only 10% of the free sulphide is HS⁻. This suggests why *R. pachyptila* is observed in higher pH waters with low Fe_T and AVS concentrations, as compared with *A. pompejana*. Where the pH in diffuse flow waters is typically higher than 6.5 and iron is present in micromolar or lower amounts, as at Guaymas Basin in the Gulf of California, *R. pachyptila* is the dominant macrofaunal form and *A. pompejana* is not found. The presence of H₂S along with the two endosymbiont-containing vent bivalves (*Bathymodilus thermophilus* and *Calyptogena magnifica*), which also inhabit low-temperature diffuse flow areas, follows a similar pattern to that of *R. pachyptila*.

In situ electrochemical speciation and discrete sample measurements (temperature, pH, sulphide, Fe_T) provided unprecedented correlations with the distribution of certain hydrothermal vent

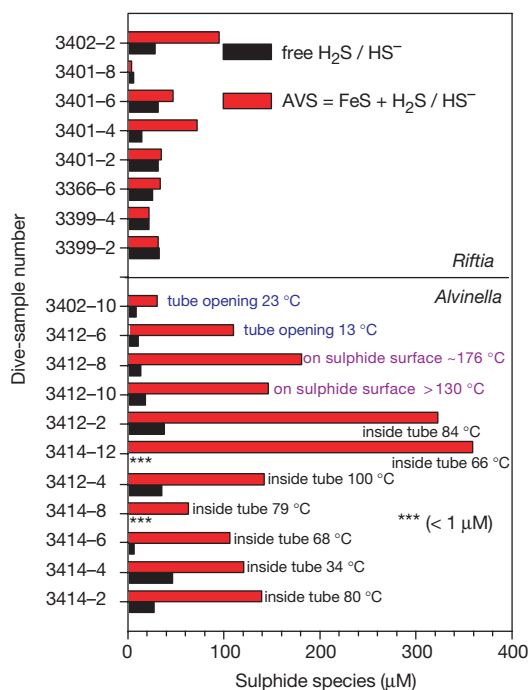


Figure 2 Histogram of free H₂S/HS⁻ and AVS data from discrete samples. For *Alvinella* samples at temperatures higher than 40 °C, temperature fluctuated rapidly during the *in situ* voltammetry scans, therefore free H₂S/HS⁻ data are from samples analysed on board ship. Figure 1d is from sample 3414-2 and shows no free H₂S/HS⁻; thus, FeS

dissociates to Fe²⁺ and free H₂S/HS⁻ on cooling and transport of samples to the surface. Free H₂S/HS⁻ data plotted for *Alvinella* samples are an overestimate of the actual free H₂S/HS⁻ that the organisms actually experience. For *Riftia* samples, both *in situ* and onboard voltammetry data give similar free H₂S/HS⁻ data.

macrofauna and evidence for the close association of vent community development with physiochemical conditions. *R. pachyptila* is found in modest lower temperature environments with low Fe_T , circumneutral pH and sufficient quantities of free sulphide to support robust bacterial endosymbiosis. The ability of *A. pompejana* to thrive in extreme microhabitats is made possible by the unique interaction of Fe^{2+} and free sulphide to form FeS species at high temperature, which lowers pH and detoxifies sulphide by rendering it biologically unavailable. Our *in situ* analyses have demonstrated the first real opportunity, at biologically relevant spatial scales, to show how differences in chemical speciation control the unique ecology of vent environments. The interplay of O_2 , H_2S , Fe^{2+} and FeS (aq) in controlling biology in primordial environments on Earth could provide a paradigm for the detection of life on other planets. □

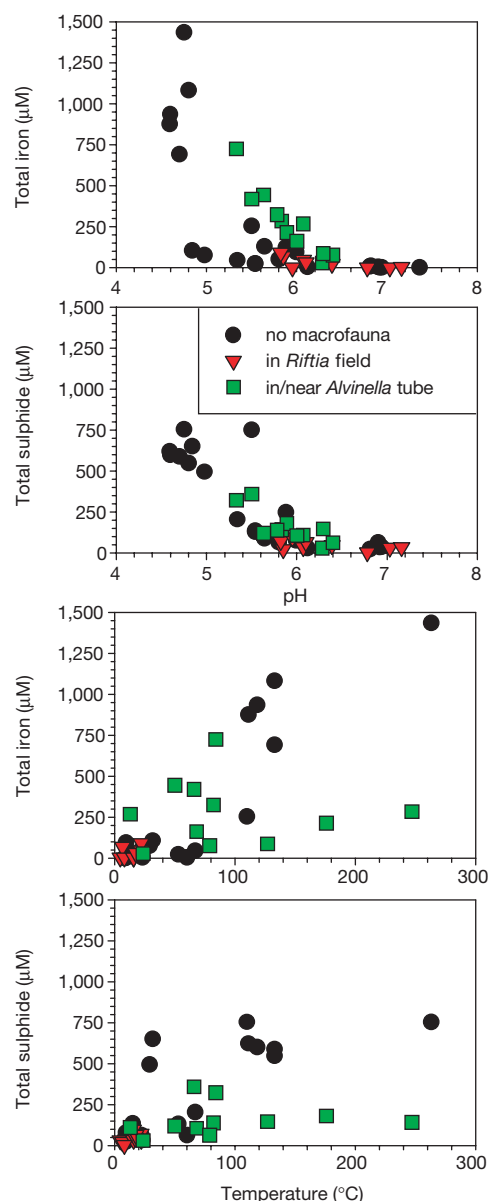


Figure 3 Fe/AVS versus pH/temperature plots from discrete samples. Circles, samples where no macrofauna exist; triangles, samples from *Riftia* sites; squares, samples from in or near *Alvinella* tubes. Correlations for *Alvinella*: Fe versus pH ($r^2 = 0.880$); Fe versus temperature ($r^2 = 0.0003$); AVS versus pH ($r^2 = 0.647$); AVS versus temperature ($r^2 = 0.018$). For *Riftia*: Fe versus pH ($r^2 = 0.533$); Fe versus temperature ($r^2 = 0.156$); AVS versus pH ($r^2 = 0.280$); AVS versus temperature ($r^2 = 0.327$).

Methods

Electrochemical equipment

The *in situ* analyser consists of a pressure case enclosing the voltammetry hardware linked to an IBM compatible computer. This computer communicates with another in the *Alvin* through a 15-m RS 232 cable and is controlled by an operator, who can reprogram waveforms to respond to the radically different environments found at vents. A separate 1-m cable was used to make connections with the working, counter, reference electrodes and the pressure housing. This cable had four inputs for working electrodes, which were controlled by a multiplexer, one input for the counter electrode and another for the reference electrode. Another input for grounding the reference electrode from the submersible ensured signal integrity. The working electrodes are solid-state gold-amalgam (Au/Hg), and were prepared in polyethylene terephthalate (PEEK) tubing¹⁴. The reference electrode was Ag wire oxidized in 3 M KCl to make an AgCl coating. This Ag/AgCl reference was used as a solid-state electrode as seawater has a consistent ionic strength of 0.7. Potentials for chemical species measured *in situ* and on board ship with the same reference electrode were comparable so no pressure effects were observed. Scan rates were 1,000 mV s⁻¹ unless stated otherwise. Conditioning between scans was performed at -0.9 V, where none of the chemical species is electroactive^{13,14}. Potential scan direction was from positive to negative for LSV. For cyclic voltammograms, the potential scan was positive to negative back to positive.

Analyses

The electrodes were mated with a thermocouple and syringe sampler ('sipper') to obtain discrete samples for shipboard analyses²⁰. In some experiments, the electrodes, thermocouple and sipper tube were coupled as a sensor package for direct insertion into *Alvinella* tubes to analyse that chemical environment directly. We measured pH by potentiometry on these samples immediately after bringing them to the surface. Aliquots of each sample were preserved in basic zinc acetate solution for subsequent AVS analysis by square-wave voltammetry on a hanging-drop mercury electrode after a 3-M HCl acid leach to release H_2S , which was trapped in 1 M NaOH (ref. 23). This method gave identical results with the methylene blue method¹⁵. Fe_T was measured on subsamples stored in nitric acid²⁴. We also measured unfiltered samples by cyclic voltammograms on board ship to confirm any chemical speciation changes.

Received 22 November 2000; accepted 16 January 2001.

- Cavanaugh, C. et al. Prokaryotic cells in the hydrothermal vent tube worm *Riftia pachyptila*: Possible chemosymbiotic symbionts. *Science* **213**, 340–342 (1981).
- Felbeck, H. Chemosymbiotic potential of the hydrothermal vent tube worm *Riftia pachyptila* Jones (Vestimentifera). *Science* **213**, 336–338 (1981).
- Childress, J. J. & Mickel, T. J. Oxygen and sulfide consumption rates of the clam *Calyptogena pacifica*. *Mar. Biol. Lett.* **3**, 73 (1982).
- Goffredi, S. K., Childress, J. J., Desaulniers, N. T. & Lallier, F. H. Sulfide acquisition by the vent worm *Riftia pachyptila* appears to be via uptake of HS^- , rather than H_2S . *J. Exp. Biol.* **200**, 2609–2616 (1997).
- Desbruyères, D. et al. Biology and ecology of the 'pompeii worm' (*Alvinella pompejana* Desbruyères and Laubier) a normal dweller of an extreme deep-sea environment: a synthesis of current knowledge and recent developments. *Deep-Sea Res. II* **45**, 383–422 (1998).
- Sarradin, P. M. et al. Chemical and thermal description of the environment of the genesis hydrothermal vent community (13°N, EPR). *Cah. Biol. Mar.* **39**, 159–167 (1998).
- Shank, T. M. et al. Temporal and spatial patterns of biological community development at nascent deep-sea hydrothermal vents (9°50'N, East Pacific Rise). *Deep-Sea Res. II* **45**, 465–515 (1998).
- Cary, S. C., Shank, T. & Stein, J. Worms bask in extreme temperatures. *Nature* **391**, 545–546 (1998).
- Chevaldonné, P., Desbruyères, D. & Childress, J. J. Some like it hot and some even hotter. *Nature* **359**, 593–594 (1992).
- Juniper, S. K. & Martineau, P. Alvinellids and sulfides at hydrothermal vents of the eastern Pacific: a review. *Am. Zool.* **35**, 74–185 (1995).
- Johnson, K. S., Beehler, C. L., Sakamoto-Arnold, C. M. & Childress, J. J. In situ measurements of chemical distributions in a deep-sea hydrothermal vent field. *Science* **231**, 1139–1141 (1986).
- Von Damm, K. L. in *Seafloor Hydrothermal Systems* (eds Humphris, S. E., Zierenberg, R. A., Mullineaux, L. S. & Thomson, R. E.) 222–247 (American Geophysical Union, Washington DC, 1995).
- Brendel, P. J. & Luther, G. W. III Development of a gold amalgam voltammetric microelectrode for the determination of dissolved Fe, Mn, O₂ and S(-II) in porewaters of marine and freshwater sediments. *Environ. Sci. Technol.* **29**, 751–761 (1995).
- Luther, G. W. III, Reimers, C. E., Nuzzio, D. B. & Lovatolo, D. In situ deployment of voltammetric, potentiometric and amperometric microelectrodes from a ROV to determine O₂, Mn, Fe, S(-2) and pH in porewaters. *Environ. Sci. Technol.* **33**, 4352–4356 (1999).
- Cline, J. E. Spectrophotometric determination of hydrogen sulfide in natural waters. *Limnol. Oceanogr.* **14**, 454–458 (1969).
- Rickard, D. T. & Luther, G. W. III Kinetics of pyrite formation by the H_2S oxidation of iron(II) monosulfide in aqueous solutions between 25 °C and 125 °C: the mechanism. *Geochim. Cosmochim. Acta* **61**, 135–147 (1997).
- Theberge, S. M. & Luther, G. W. III Determination of the electrochemical properties of a soluble aqueous FeS cluster present in sulfidic systems. *Aq. Geochem.* **3**, 191–211 (1997).
- Rozan, T. E., Theberge, S. M. & Luther, G. W. III Quantifying elemental sulfur (S⁰), bisulfide (HS⁻) and polysulfides (S_n²⁻) using a voltammetric method. *Anal. Chim. Acta* **415**, 175–184 (2000).
- Cary, S. C., Cottrell, M. T., Stein, J. L., Camacho, F. & Desbruyères, D. Molecular identification and localization of filamentous symbiotic bacteria associated with the hydrothermal vent annelid *Alvinella pompejana*. *Appl. Environ. Microbiol.* **63**, 1124–1130 (1997).
- Di Meo, C. A., Wakefield, J. R. & Cary, S. C. A new device for sampling small volumes of water from marine micro-environments. *Deep-Sea Res. I* **46**, 1279–1287 (1999).
- Rickard, D. T. Kinetics of pyrite formation by the H_2S oxidation of iron(II) monosulfide in aqueous solutions between 25 °C and 125 °C: the rate equation. *Geochim. Cosmochim. Acta* **61**, 115–134 (1997).
- Millero, F. J., Plese, T. & Fernandez, M. The dissociation of hydrogen sulfide in seawater. *Limnol. Oceanogr.* **33**, 269–274 (1988).

23. Luther, G. W. III, Ferdelman, T. G., Kostka, J. E., Tsamakis, E. J. & Church, T. M. Temporal and spatial variability of reduced sulfur species (FeS_2 , $\text{S}_2\text{O}_3^{2-}$) and porewater parameters in salt marsh sediments. *Biogeochemistry* **14**, 57–88 (1991).
24. Stookey, L. L. Ferrozine—a new spectrophotometric reagent for iron. *Anal. Chem.* **41**, 779–782 (1970).

Acknowledgements

We thank A. L. Reysenbach, K. Longnecker, the DSV *Alvin* pilots (P. Hickey, S. Faluotico and B. Williams), and the crew and Captain of the R/V *Atlantis* for their help and encouragement. This work was funded by grants from the National Science Foundation and the National Aeronautics and Space Administration, and by the University of Delaware Sea Grant Program through the National Oceanic and Atmospheric Administration, and the Devonshire Foundation.

Correspondence and requests for materials should be addressed to G.W.L. (e-mail: luther@udel.edu).

Perceiving visual expansion without optic flow

Paul R. Schrater^{*†}, David C. Knill^{†‡} & Eero P. Simoncelli[§]

^{*} Department of Neuroscience, University of Pennsylvania, 215 Stemmler Hall, Philadelphia, Pennsylvania 19104, USA

[‡] Department of Psychology, University of Pennsylvania, 3815 Walnut Street, Philadelphia, Pennsylvania 19104, USA

[§] Howard Hughes Medical Institute and Center for Neural Science and Mathematics, New York University, 4 Washington Place, Rm 809, New York, New York 10003, USA

When an observer moves forward in the environment, the image on his or her retina expands. The rate of this expansion conveys information about the observer's speed¹ and the time to collision^{2–4}. Psychophysical^{5–7} and physiological^{8,9} studies have provided abundant evidence that these expansionary motions are processed by specialized mechanisms in mammalian visual systems. It is commonly assumed that the rate of expansion is estimated from the divergence of the optic-flow field (the two-dimensional field of local translational velocities)^{10–14}. But this rate might also be estimated from changes in the size (or scale) of image features¹⁵. To determine whether human vision uses such scale-change information, we have synthesized stochastic texture stimuli in which the scale of image elements increases gradually over time, while the optic-flow pattern is random. Here we show, using these stimuli, that observers can estimate expansion rates from scale-change information alone, and that pure scale changes can produce motion after-effects. These two findings suggest that the visual system contains mechanisms that are explicitly sensitive to changes in scale.

To test the hypothesis that the rate of image expansion on the retina might be estimated from changes in scale (Fig. 1), we designed three types of stochastic expansion stimuli (Fig. 2a). We created a sequence of uncorrelated white-noise images, and convolved each one with a band-pass filter whose centre frequency decreases exponentially with time. The result is a sequence of random textures whose average spatial scale grows according to a fixed rate of expansion. Because the individual stimulus frames are uncorrelated, both the optic-flow field and its divergence are (on average) zero, and thus cannot be used to compute expansion rate. Nevertheless, most people who have viewed these stochastic stimuli report a strong perception of expansion, emanating from the point of fixation. This suggests that humans have mechanisms that are

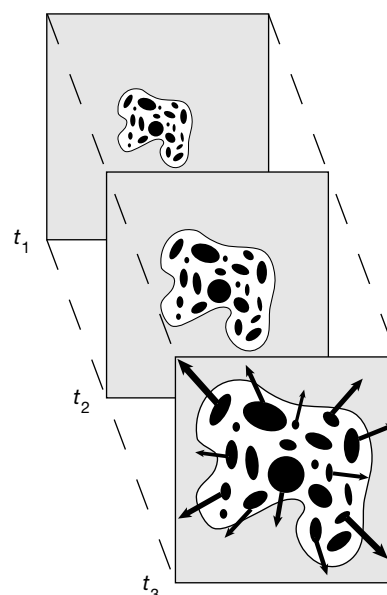


Figure 1 Illustration of visual input when an observer moves forward in the environment. Shown are three frames taken from a sequence of visual images (a movie). Superimposed on the last frame are example optic-flow vectors, indicating the local translational motion of selected points in the scene. The radiating pattern of these vectors indicates that the observer is moving forward. The divergence of this vector field provides one source of information about the rate of expansion. The change in scale of the objects or texture elements provides another source of information.

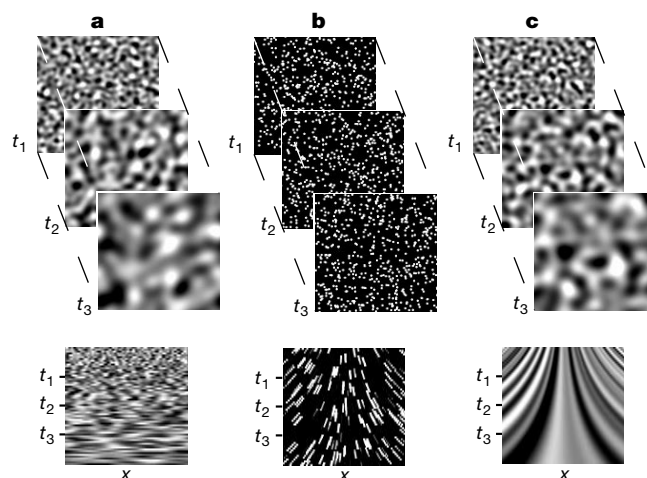


Figure 2 Illustration of the three types of stimuli used in our experiments. Top, example x - y frames of the stimulus movie; bottom, x - t cross-sections. **a**, Stochastic texture stimulus. Each frame contains a random texture synthesized by convolving a gaussian white-noise image with a band-pass filter. The peak spatial frequency of the band-pass filter decreases over time, producing a gradual change in scale of texture elements. The x - t slice illustrates the lack of consistent displacement information. Because each frame is generated from independent white noise, the resulting optic-flow field is random, and thus cannot be used to estimate expansion rate. **b**, Limited-lifetime random dots stimulus. Dots appear asynchronously in random locations, move according to the global expansion pattern for five frames, and then disappear. For this stimulus, the optic-flow field may be used to estimate expansion rate, but there is minimal scale-change information. **c**, Deterministic texture stimulus. Each frame is a re-scaled replica of a single image of band-pass-filtered gaussian white noise. Both optic flow and scale changes provide information about the expansion rate.

[†] Present address: Department of Psychology, University of Minnesota, 75 E. River Road, Minneapolis, Minnesota 55455, USA (P.R.S.); Center for Visual Science, University of Rochester, Rochester, New York 14627, USA (D.C.K.).

sensitive to changes in scale per se.

To test this hypothesis more rigorously, we measured the ability of subjects to match the expansion rate of our stochastic stimuli to the expansion rate of a field of dots undergoing constant expansionary motion. The dots in the latter stimuli have a limited lifetime (Fig. 2b; and Methods), and thus contain optic-flow information about the rate of expansion but minimal scale-change information. On each trial, observers fixated on a point lying between a stochastic expansion stimulus and an expanding dots stimulus. The stimulus movies were repeated while the observer adjusted the rate of expansion of the dots until it appeared to match the expansion rate of the stochastic stimulus. We measured the matching expansion rates for four different stimulus conditions: two durations and two average texture scales. To provide a standard against which to compare these matches, observers were also asked to perform the same task using deterministic texture expansion movies with identical parameters (see Fig. 2c).

Figure 3 shows the results of the matching experiment for both stochastic and deterministic texture expansion stimuli. All obser-

vers show consistent increases in matched dot expansion rates as a function of the expansion rate simulated by the texture stimuli. Although matches are sometimes far from veridical, they are no more so for the stochastic stimuli than for the deterministic stimuli. Linear regression on the matching data shows some variation in slope and intercept across the four different stimulus conditions, but no consistent pattern emerges across subjects.

Although individual subjects show differences in the slopes of matching functions between stochastic and deterministic stimuli, group averages show no significant difference in the average slope for the two types of stimuli (analysis of variance (ANOVA): $F_{1,38} = 0.49$, $P = 0.49$, mean slope = 0.91 and 0.95, respectively). In addition, group averages of the r^2 values of the fits are not significantly different (t -test: $t_{38} = 1.6$, $P = 0.11$, mean $r^2 = 0.44$ and 0.52, respectively). Note, however, that subjects' matches for the deterministic stimuli do have a lower variance than those for the stochastic stimuli (mean s.d. = 0.53 and 0.65; $F_{1,158} = 12.7$, $P < 0.001$). This implies that observers can use the optic-flow information available in the deterministic expansion stimuli to

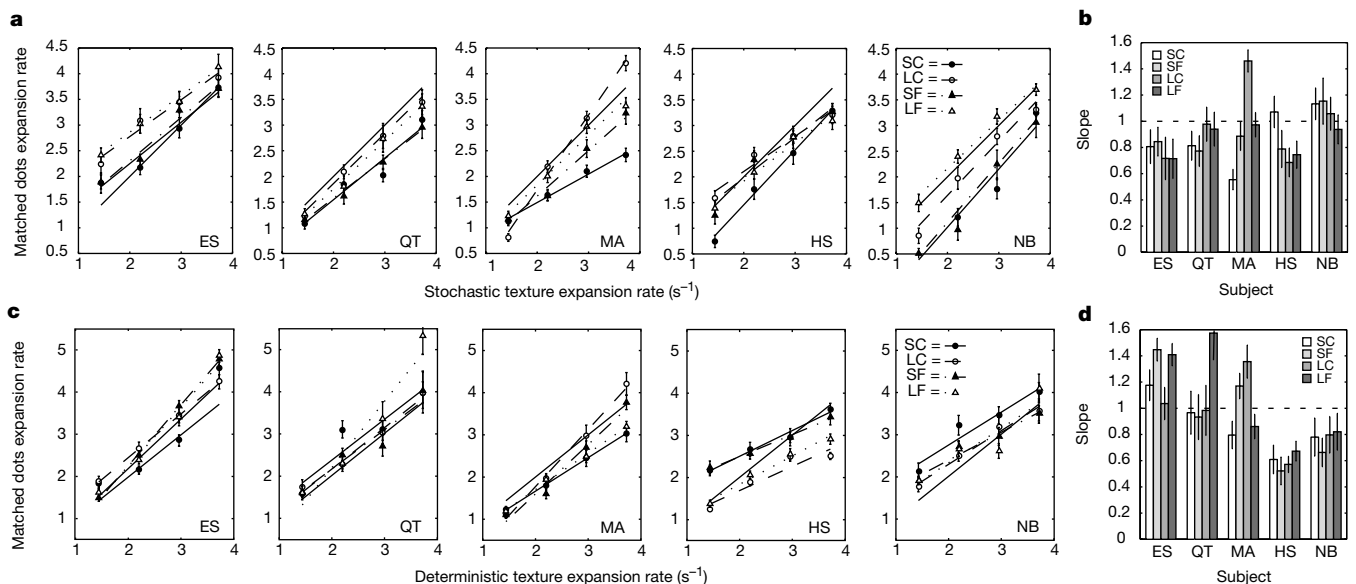


Figure 3 Perceptual matches of expansion rate between dot and texture stimuli.

a, Perceptually matched expansion rate of dot stimuli as a function of the stochastic texture expansion rate for five observers. Symbols indicate mean of 20–30 trials; bars indicate standard error. Stimulus conditions are denoted by different plot symbols: black, short duration (0.43 s); white, long duration (0.73 s); circles, coarse scale (geometric mean spatial frequency = 0.9 cycles per degree); triangles, fine scale (2.1 cycles per

degree). Thin lines indicate least-squares linear fits for each condition; thick line indicates a veridical match (unit slope). **b**, Slope estimates from **a**. Error bars indicate 95% confidence intervals. Dashed horizontal line denotes a veridical match. **c**, Perceptual matches between expansion rates of deterministic texture stimuli and dot stimuli (details as in **a**). **d**, Slope estimates from **c**.

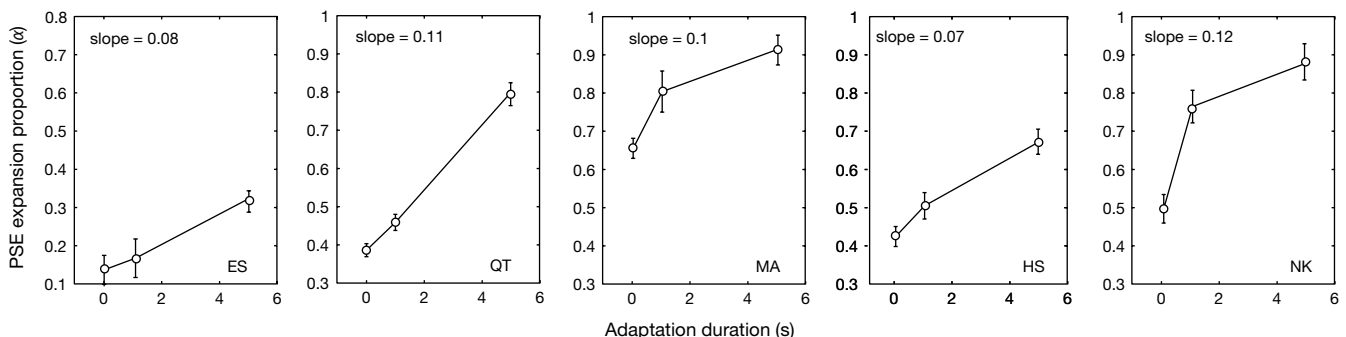


Figure 4 Effect of adapting to stochastic expansion stimuli on the perception of an ambiguous test stimulus for five subjects. The value of α yielding 50% expansion/contraction judgments is plotted against adaptation duration (zero corresponds to no

adaptation). Error bars indicate standard errors. Also indicated is the slope of the psychometric function that was fit to the raw data, which was constant across different adaptation conditions.

improve the reliability of their matches.

The matching experiment leaves open the possibility that subjects use a cognitive strategy to estimate the expansion rates of stochastic expansion stimuli (for example, by estimating the static texture size at two or more moments in time, and dividing by the time interval between them). To rule this out, we performed an additional experiment to measure motion after-effects induced by the stochastic stimuli, as these effects are generally believed to be insensitive to cognitive intrusion^{16,17}. We used an experimental design analogous to one used previously to measure adaptation to fields of coherent translating dots¹⁸. In these experiments, adaptation stimuli are followed by ambiguously moving test stimuli containing mixtures of leftward and rightward moving dots. Adaptation to rightward motion causes a balanced mixture to appear to move leftward. The strength of this after-effect is measured as the percentage of test stimulus rightward dots required to produce a perception of no net motion.

Our experiment used stochastic expansion textures as adaptation stimuli. Test stimuli were created by mixing frames from expanding and contracting stochastic texture movies. In particular, we first created a set of stochastic expansion movies with frequency and duration parameters identical to those of the adaptation stimuli. We converted half of these into contraction movies by reversing the frame order. Ambiguous test stimuli were created by selecting each frame at random from either an expansion or contraction movie.

A mixture parameter, α , controlled the proportion of frames drawn from expansion movies (see Methods). Thus, $\alpha = 1$ corresponds to a stimulus whose frames are all drawn from expansion movies. On each trial, subjects viewed the adaptation stimulus for 0 (no adaptation), 1 or 5 s, followed by a test stimulus, after which they indicated whether the test stimulus appeared to be expanding or contracting. The α -value used to create the test stimuli was adjusted using a psychometric procedure to produce expanding responses 20, 50 and 80% of the time. The resulting data were fitted with a cumulative Normal function with a constant slope across adaptation conditions (nested model comparison tests did not support the use of multiple slopes for any of the subjects: $\chi^2(2) < 0.2$ for all subjects).

Figure 4 shows the null (50%) points estimated for each of five subjects as a function of adaptation time. With no adaptation, three out of five observers perceived the ambiguous stimuli as expanding, consistent with previously reported biases for forward/backward motion perception¹⁹. As the adaptation time increases, subjects' expansion null points increase, indicating that adaptation to stochastic expansion stimuli produces a bias toward perceiving the test stimuli as contracting. After 5 s of adaptation, the bias is substantial: stimuli that would have been judged ordinarily to be expanding 99% of the time became completely ambiguous. In addition, as the effect increases with adaptation duration, the shifts must be caused by a real after-effect rather than an induction or priming effect.

The results of our matching and adaptation experiments provide compelling evidence for the hypothesis that the human visual system uses scale-change information to compute expansion rate¹⁵. The strength of the evidence is due to the stochastic texture stimuli, which neatly separate the optic-flow information from the scale-change information. Related stereoscopic stimuli have been used to show that the perception of surface slant from stereopsis does not require point-wise disparity estimates²⁰, and related sound signals have been used to show neural sensitivity to frequency modulations in the mammalian auditory system²¹.

The mechanism by which scale-change information is processed, and the method by which it is combined with optic-flow information remain unknown. Two basic possibilities exist. The visual system might use a single mechanism that is sensitive to both scale-change and optic-flow information. Alternatively, the visual system might have separate mechanisms for computing expansion from scale-change and optic-flow information, combining the

outputs of the two to arrive at a unitary estimate. For example, expansion can be computed separately from scale-change information by estimating the 'motion' of the signal's Fourier energy pattern along the spatial frequency axis. Given a set of neurons whose responses approximate local Fourier energy in a frequency band (like V1 complex cells), this computation could be neurally implemented as time-varying pooling of these neurons' responses adjusted at each moment to track the expected scale of the expanding image. These issues might be addressed best by physiological and psychophysical experiments using the stimuli introduced in this paper. $\square$

Methods

Stimuli

All stimuli were displayed at 60 Hz in 8 bits of precision on a calibrated gray-scale monitor. Stimuli were viewed monocularly from a distance of 55.7 cm, and subtended 16.5 degrees of visual angle. Stochastic textures were constructed by convolving independent frames of gaussian white noise with band-pass filters that are described in the spatial frequency domain by: $F(\omega_s) = 0.5 + 0.5 \cos[\omega_s/2f(t)]$, where $\cos$ is a cosine function restricted to one cycle, $f(t)$ is the peak frequency of the filter at time t , and $\omega_s = \sqrt{(\omega_x^2 + \omega_y^2)}$.

For a constant expansion rate a , $f(t) = \exp(\log[f_{\text{initial}}] - at)$. We used four different expansion rates: 0.96, 1.44, 2.0 or 2.75 s^{-1} . For each rate, we used two different durations (0.43 and 0.73 s) and two geometric-mean frequencies (0.9 and 2.1 cycles per degree). For a given geometric mean frequency f_{mean} , the initial and final frequencies (f_{initial} and f_{final}) were defined by $\log[f_{\text{initial,final}}] = aT/2 \pm \log[f_{\text{mean}}]$, where T is the duration. We constructed 125 different versions of each stochastic stimulus type by interleaving every third frame of 5 different movies. The deterministic stimuli were made from the stochastic texture stimuli by taking the first frame and performing band-limited sub-sampling and interpolation to digitally expand the initial image. For the dots stimuli, 200 constant size (4 pixels), randomly placed dots were moved according to: $x(t+1) = x(t) + ax(t)$, which approximates $dx/dt = ax$. Dots had limited lifetimes of five frames (83 ms), and expired asynchronously. When the lifetime was exceeded, or when dots advanced past the edge of the image, they were randomly re-placed in the image. This procedure produces a constant expected density of dots across the image, and prevents the dots from carrying significant size-change information.

Matching experiment

Observers were asked to match the perceived expansion rate of the stochastic textures with that of the dots. Observers were given no feedback, and were queried at the end of the experiment regarding the phenomenal appearance of the stimuli and their matching strategy. Each trial consisted of the following events. Observers fixated on a spot between a stochastic texture movie centred 9.1 degrees left, and a dots movie centred 9.1 degrees right of fixation. Both movies were repeated with a 0.33-s delay between cycles. At any time observers could press keys to increase or decrease the expansion rate of the dots movie, or indicate that the expansion rates appeared to match. Between 24 and 34 matches were collected for each of the 16 different stochastic texture movies, the first 4 of which were discarded to remove early learning effects. The observer's matches for each condition were fit by least-squares linear regression.

Adaptation experiment

Ambiguous test stimuli for the adaptation experiment were constructed by sampling frames from stochastic texture expansion movies, such that the n th frame of the test movie is copied from the k th frame of an expanding stochastic texture movie according to the distribution: $p(k|n, \alpha) = \alpha\phi(n) + (1 - \alpha)\phi(N - n + 1)$, where $\phi(n)$ denotes a Normal density function with mean n and variance 1.5, rounded to the nearest integer, and N is the total number of frames in the movie. Observers viewed these stimuli centrally and after each presentation chose whether the stimuli appeared to be expanding or contracting. The QUEST adaptive psychophysical procedure²² was used to estimate the α producing 20, 50 and 80% expansion responses. The test stimuli were preceded by adaptation stimuli of three different durations (0, 1 and 5 s). The 5-s adaptation condition also included an initial 20-s adaptation period. For each adaptation condition, 400 trials were collected and all response data were fit by a cumulative Normal psychometric function through maximum likelihood. Fifty per cent null points were determined from these psychometric function fits, and standard errors of these estimates were estimated using 1,000 fits to bootstrap re-samples of the response data. Differences in slope across adaptation duration were assessed using a nested-hypothesis test procedure²³. The test determines whether the goodness-of-fit of a model with independent slopes for each adaptation duration is significantly better than a single slope model by computing $x = 2 \ln[L_i/L_r]$, where L_i and L_r are the likelihoods of the fit for the independent and reduced models, respectively. The statistic x is χ^2 distributed with 2 degrees of freedom, computed from the difference in the number of free parameters in the two models.

Received 3 October 2000; accepted 20 January 2001.

1. Clifford, C. W., Beardsley, S. A. & Vaina, L. M. The perception and discrimination of speed in complex motion. *Vision Res.* **39**, 2213–2227 (1999).
2. Lee, D. N. A theory of visual control of braking based on information about time-to-collision. *Perception* **5**, 437–459 (1976).
3. Schiff, W. & Detwiler, M. L. Information used in judging impending collision. *Perception* **8**, 647–658 (1979).

4. Kaiser, M. K. & Hecht, H. Time-to-passage judgments in non-constant optical flow fields. *Percept. Psychophys.* **57**, 817–825 (1995).
5. Regan, D. & Beverley, K. I. Illusory motion in depth: aftereffect of adaptation to changing size. *Vision Res.* **18**, 209–212 (1978).
6. Regan, D. & Beverley, K. I. Binocular and monocular stimuli for motion in depth: changing-disparity and changing-size feed the same motion-in-depth stage. *Vision Res.* **19**, 1331–1340 (1979).
7. Morrone, M. C., Burr, D. C., Di Pietro, S. & Stefanelli, M. Cardinal directions for visual optic flow. *Curr. Biol.* **9**, 763–766 (1999).
8. Orban, G. A. *et al.* First-order analysis of optical flow in monkey brain. *Proc. Natl Acad. Sci. USA* **89**, 2595–2599 (1992).
9. Duffy, C. J. & Wurtz, R. H. Sensitivity of MST neurons to optic flow stimuli: I. a continuum of response selectivity to large-field stimuli. *J. Neurophysiol.* **65**, 1329–1345 (1991).
10. Gibson, J. J. *The Perception of the Visual World* (Houghton Mifflin, Boston, 1950).
11. Warren, W. H., Morris, M. W. & Kalish, M. L. Perception of translational heading from optical flow. *J. Exp. Psychol. Hum. Percept. Perform.* **14**, 646–660 (1988).
12. Warren, W. H., Blackwell, A. W., Kurtz, K. J., Hatsopoulos, N. G. & Kalish, M. L. On the sufficiency of the velocity field for perception of heading. *Biol. Cybern.* **65**, 311–320 (1991).
13. Perrone, J. A. Model for the computation of self-motion in biological systems. *J. Opt. Soc. Am. A* **9**, 177–194 (1992).
14. Crowell, J. A. & Banks, M. S. Perceiving heading with different retinal regions and types of optic flow. *Percept. Psychophys.* **53**, 325–337 (1993).
15. Beverley, K. I. & Regan, D. Texture changes versus size changes as stimuli for motion in depth. *Vision Res.* **23**, 1387–1400 (1983).
16. Anstis, S. M. The perception of apparent movement. *Phil. Trans. R. Soc. Lond. B* **290**, 153–168 (1980).
17. Cavanagh, P. & Mather, G. Motion: the long and short of it. *Spatial Vis.* **4**, 103–129 (1989).
18. Blake, R. & Hiris, E. Another means for measuring the motion aftereffect. *Vision Res.* **33**, 1589–1592 (1993).
19. Perrone, J. A. Anisotropic responses to motion toward and away from the eye. *Percept. Psychophys.* **39**, 1–8 (1986).
20. Tyler, C. & Sutter, E. Depth from spatial frequency difference: An old kind of stereopsis? *Vision Res.* **19**, 859–865 (1979).
21. Tian, R. & Rauschecker, J. P. Processing of frequency-modulated sounds in the cat's posterior auditory field. *J. Neurophysiol.* **79**, 2629–2642 (1998).
22. Watson, A. B. & Pelli, D. G. Quest: A Bayesian adaptive psychophysical method. *Percept. Psychophys.* **33**, 113–120 (1983).
23. Kendall, M. K. & Stuart, A. *The Advanced Theory of Statistics: Vol. 3. Design, Analysis, and Time Series* (Hafner, New York, 1966).

Acknowledgements

This work was begun when all three authors were at the University of Pennsylvania. P.S. was supported by a training grant from the NEI at the University of Pennsylvania and by a research grant from the NIH at the University of Minnesota; E.S. was supported by the National Science Foundation and the Alred P. Sloan Foundation; and D.K. was supported by a research grant from the NIH.

Correspondence and requests for materials should be addressed to P.S. (e-mail: schrater@eye.psych.umn.edu).

The motor side of depth vision

Kai Schreiber^{*†‡}, J. Douglas Crawford^{†‡}, Michael Fetter[§] & Douglas Tweed^{*†‡}

^{*} Departments of Physiology and Medicine, University of Toronto, 1 King's College Circle, M5S 1A8 Toronto, Canada

[†] Canadian Institutes of Health Research, Group for Action and Perception

[‡] Centre for Vision Research, York University, 4700 Keele Street, M3J 1P3 Toronto, Canada

[§] Department of Neurology, University Hospital Tübingen, Hoppe-Seyler-Straße 3, 72072 Tübingen, Germany

To achieve stereoscopic vision, the brain must search for corresponding image features on the two retinas¹. As long as the eyes stay still, corresponding features are confined to narrow bands called epipolar lines^{2,3}. But when the eyes change position, the epipolar lines migrate on the retinas^{4–6}. To find the matching features, the brain must either search different retinal bands depending on current eye position, or search retina-fixed zones that are large enough to cover all usual locations of the epipolar lines. Here we show, using a new type of stereogram in which the depth image vanishes at certain gaze elevations, that the search zones are retina-fixed. This being the case, motor control acquires a crucial function in depth vision: we show that the eyes twist

about their lines of sight in a way that reduces the motion of the epipolar lines, allowing stereopsis to get by with smaller search zones and thereby lightening its computational load.

Eye motion shifts the epipolar lines—the retinal bands where corresponding image features, or stereo matches, project in the two eyes. In the simulation shown in Fig. 1, if the small open circles represent point images cast onto the right retina, and the eyes are converged 30° while looking level, then the corresponding images on the left retina must lie somewhere along the thick lines. When the eyes look 30° down, still converged 30°, the images on the left retina must lie on the thin lines in Fig. 1. When the eyes look 30° up, the images lie on the dashed lines. If the brain searches the wrong lines for stereo matches, it will not find them and stereopsis will fail.

This is a fundamental problem facing any creature that wants both stereopsis and mobile eyes, and there are just two possible solutions. Either the matching algorithm searches different retinal bands depending on current eye position, or it searches eye-fixed zones large enough to encompass all possible locations of the epipolar lines in all usual eye configurations (outlined regions in Fig. 1). Most theories of stereopsis assume that the brain searches narrow bands, so as to reduce its computational load. We used visual stimuli like those in Fig. 2 to show that in fact the brain uses large, eye-fixed zones.

Discs a1 and a2 in Fig. 2 form a cyclorotated stereogram (b1 and b2 form another, interleaved to save space on the page). Disc a1, to be viewed by the right eye, is rotated 2° counterclockwise, and a2 is rotated 2° clockwise, for a total of 4° incyclorotation. When humans converge and look down, our eyes incycloverge, with the upper poles of both eyes rotating medially^{5,7,8}. When we converge and look up, our eyes excycloverge. If stereo matching adjusts its search to correct for eye position, this cyclovergence should make no difference to vision; but if the matching algorithm does not correct for eye position, we should see incyclorotated stereograms better when we look down, and excyclorotated stereograms better when we look up. If you fuse the stereograms of Fig. 2 as instructed in the figure legend, you should see the stereoscopic images only on downgaze.

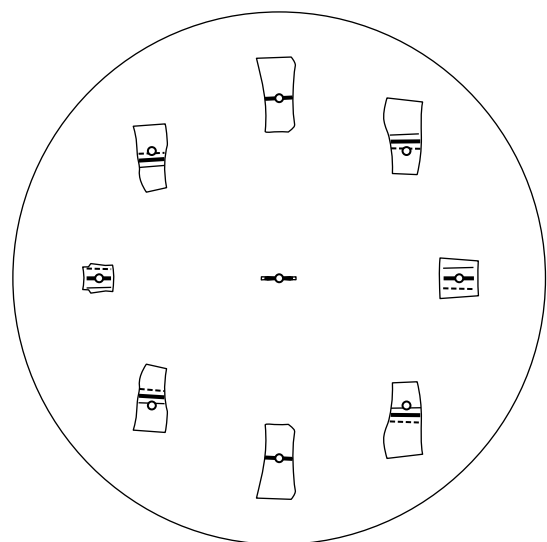


Figure 1 Epipolar lines vary with eye position. The nine open circles are images of nine objects projected onto the right retina, one foveal and the rest 15° eccentric (the perimeter circle marks 22.5° from the fovea). Projections of these same objects onto the left retina must lie somewhere on the thick line segments—pieces of epipolar lines—when the eyes converge 30° and look level. They lie on the dashed lines when the eyes look up 30° and on the thin lines when they look down 30°, always with 30° of convergence. Outlines mark the entire range of motion of the epipolar segments when the eyes also move $\pm 20^\circ$ horizontally. Eye motion in this simulation is a realistic combination of Listing's law and L2 (see Methods).

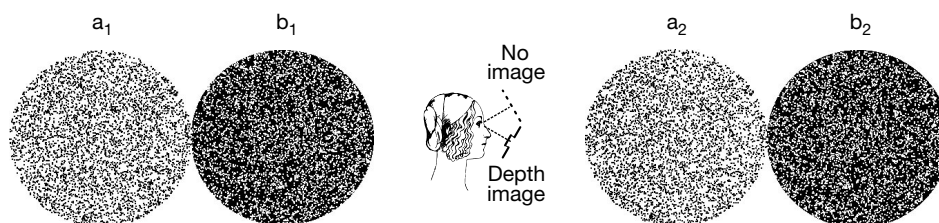


Figure 2 Cyclorotated stereograms are visible only in certain eye positions. Cross-fuse discs a1 and a2 from 30 cm away, depressing your gaze as far as possible and holding the paper orthogonal to the plane of your sight lines (not parallel to your face). You should see a depth image (a pointer, like a tapering clock hand, pointing up and right) in this position, but not when you do the same on upgaze. Flipping the page upside-down yields a

stereogram visible only on upgaze. If the image never disappears, your search zones are too large for this stereogram; try fusing discs b1 and b2, whose cyclorotation is greater (10°). If no image appears even on downgaze, your search zones are too small; try fusing a1 and a2 upside-down.

We tested five normal subjects, all of whom showed this positional stereoblindness. When one typical subject viewed stereograms while converged 30° and looking 30° down, he could reliably see depth—that is he could identify the disparity-defined image at least 50% of the time—in stereograms that were incyclorotated as much as 10.1° (Fig. 3). Looking up, his threshold was just 3.0° of incyclorotation. So a stereogram with, for example, 7° of incyclorotation was visible on downgaze but not on upgaze. For excyclorotation the subject's thresholds were better when he looked up, so that a stereogram with 7° of excyclorotation was visible on upgaze but not on downgaze. Thresholds varied between subjects, but all five showed the pattern of positional stereoblindness (Fig. 4) that one would expect if the matching algorithm were receiving no information about eye position.

Eye-position measurements quantitatively confirmed this idea. In all five subjects, we plotted cyclorotation thresholds versus cyclovergence measured by the nonius method^{4,9} (Fig. 4). If search zones are perfectly fixed to the retinas, the slope of these lines will be 1 (parallel to the dashed line in the figure), meaning that the cyclorotation thresholds rotate exactly as far as the eyes. On average, the slope was 1.06. We conclude that stereoptic search zones are essentially eye-fixed.

This finding suggests that eye control might be an important part of stereopsis. An optimized motor program could coordinate the eyes in a way that minimizes the motion of the epipolar lines and therefore allows stereopsis to get by with the smallest possible search zones, lightening the computational load. What is the optimal pattern of eye coordination? Horizontal and vertical motions of both eyes are determined by the visual fixation target, but cyclorotation is free: both eyes can turn about their lines of sight without

breaking fixation. Our simulations showed that cyclovergence—equal cyclorotation of the two eyes—makes little difference to stereopsis, but cyclovergence matters.

We considered how cyclovergence is controlled. Does it respond to visual stimuli? Do the eyes excycloverge, for example, when they see an excyclorotated stereogram? We used search coils to measure eye motion in six subjects viewing cyclorotated stereograms, and found that, on average, cyclovergence varied only 0.07° (range 0.02 – 0.16°) for every degree of stimulus cyclorotation. Larger or moving stimuli evoke stronger eye rotations^{10–13}, but even then visually driven cyclovergence is superimposed on a blind, default pattern of eye control¹⁰.

The normal pattern of eye control when viewing distant objects is Listing's law^{14,15}, which states that each eye's cyclorotation is proportional to the product of its horizontal and vertical angles in radians, $C = -HV/2$ (see Methods). By this formula, Listing's law calls for a cyclovergence range of 15.7° in our experiment, from 7.85° excyclorotation on upgaze to 7.85° incyclorotation on downgaze. But in fact, the ranges of our subjects averaged just 6.8° (Fig. 4); that is, our subjects, like those of other studies^{7,8,16–19}, violated Listing's law on near gaze, reducing their cyclovergence⁵.

Obedying Listing's law brings many functional advantages, improving motor efficiency and simplifying monocular visual processing^{5,14,20,21}. So there must be some even greater functional advantage, incompatible with Listing's law, to make the brain violate

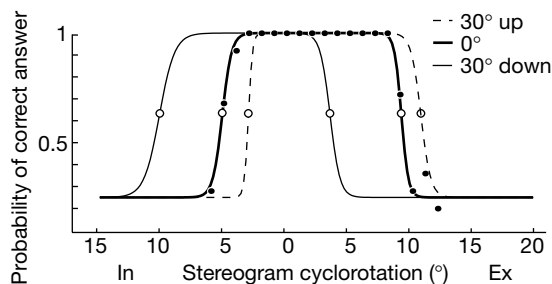


Figure 3 Stereopsis depends on gaze elevation. For this typical subject, black dots show the fraction of images that were correctly perceived or guessed—there was always a 25% chance of guessing which of the four possible disparity-defined images was present—as a function of the cyclorotation of the stereograms, when the subject converged 30° and looked level. Fitted curves plot performance at three gaze elevations. Open circles mark perception thresholds, that is, angles at which stereograms were reported correctly 62.5% of the time; this is the performance expected when the subject perceives the image 50% of the time and guesses correctly on a quarter of the remaining trials.

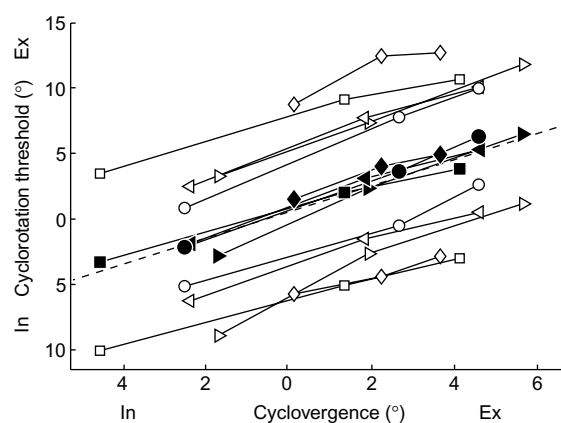


Figure 4 Perception thresholds depend on gaze elevation and cyclovergence. Open symbols at top and bottom mark excyclorotation and incyclorotation thresholds for five subjects. Filled symbols lie halfway between the two thresholds, at the midpoint of the visible range. Thresholds for all five subjects varied significantly with gaze elevation, shifting toward incyclorotation on downgaze (leftmost symbol in each string of three) and toward excyclorotation on upgaze (rightmost symbol in each string of three); $F_{2,8} = 33.40$, $P < 0.001$ for incyclorotation thresholds, $F_{2,8} = 51.95$, $P < 0.001$ for excyclorotation. These plots of perception thresholds versus cyclovergence show an average slope of 1.06, very close to slope of 1 (the dashed line) predicted if stereo matching receives no information about eye position.

it so markedly on near gaze^{5,8}. The advantage, we suggest, is that reducing cyclovergence restricts the motion of the epipolar lines, permitting stereo matching to function with smaller search zones. Still smaller zones would be possible if cyclovergence were eliminated entirely, and a further slight improvement would accrue if cyclovergence were also avoided; however, the brain rejects these motor patterns, presumably because they stray too far from Listing's law. Thus, eye control strikes a balance between the motor and monocular advantages of Listing's law and the optimization of stereoptic search.

We computed the smallest search zones that would permit stereopsis, given different patterns of eye control. Listing's law yields the black zones in Fig. 5. Eliminating cyclovergence allows stereopsis to proceed with much smaller zones—the white regions. Such plots show how the layout of an optimized visual system depends on the pattern of eye motion. As human eye control lies between these two patterns, our search zones should show an intermediate arrangement, closer to the white. (The zones may be slightly taller than predicted here because we consider only the effects of eye motion, whereas there may be other reasons for vertical search²².) As Fig. 5 shows, search zones should be flatter along the retina's horizontal meridian, taller eccentrically and smaller temporally, whereas on oblique meridians they should be vertically displaced relative to the corresponding dot on the other retina. These predictions can be tested perceptually and electrophysiologically, by plotting receptive fields of groups of visual neurons. For example, an optimal system would tolerate vertical disparities of about 0.1° within 1° of the foveal centre and 0.7° at 3° eccentricity; these predictions are consistent with the available data^{4,23,24}, and explain why vertical tolerance increases rapidly with retinal eccentricity.

Stereopsis, then, is partly a motor process. The stereo-matching algorithm does not know where it is looking, and therefore does not move its search zones on the retinas when the eyes move; instead, it uses eye-fixed search zones and relies on ocular coordination to keep corresponding images within these bounds. A motor program twists the eyes about their lines of sight in a way that reduces the motion of the epipolar lines, allowing stereopsis to operate with smaller search zones. □

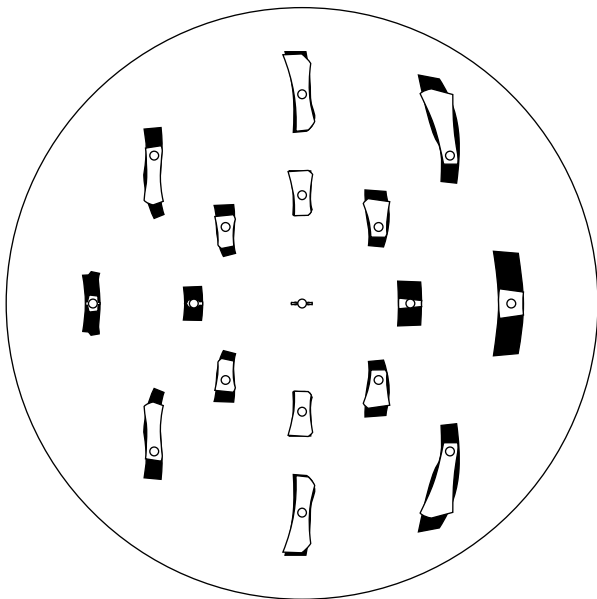


Figure 5 Optimal stereoptic search zones when the eyes obey Listing's law (black areas) and when they move without cyclovergence (white areas) in a pattern called L2 (see Methods). Rings of dots are 15° and 30° from the fovea, the perimeter circle is at 45°. Simulated eye positions ranged from -20° to 20° horizontal version, -30° to 30° vertical version and 0° to 30° vergence.

Methods

Perceptual experiments

Subjects viewed three series of 200 stereograms each at eye level and 30° above and below. Stereograms appeared on a notebook-computer screen 35 cm away in darkness, spaced so that cross-fusion required 30° of vergence; at each of the three gaze elevations, the screen was rotated so that it was orthogonal to the visual plane. Each disc of each stereogram spanned 12° and comprised 2,000 white dots on a black screen. Each stereogram contained a disparity-defined shape—a pointer, like a clock hand, pointing in one of four oblique directions, presented in random order.

Stereograms were flashed for 200 ms; subjects then chose the direction in which they thought the clock hand was pointing. An interactive program varied the cyclorotation angles of the stereograms depending on the correctness of previous answers, mapping out the threshold ranges. Because subjects viewed the stereograms cross-eyed, their lines of sight hit the flat computer screen at a 15° slant. To eliminate the optical effects of this slant, we warped the random-dot patterns on the screen in a precise way so that each retinal projection was of a disc seen face-on, without slant-induced disparities, cyclorotated about the line of sight. But as slants are present in real life when one views near objects, we also performed the experiments without the de-slanting procedure.

The results were identical, all five subjects showing the same positional stereoblindness; in this version of the experiment, the average slope of the relation between perception thresholds and cyclovergence was 1.02, almost equal to the value of 1.06 in Fig. 4. In these with-slant experiments, we presented the stereograms for 8 s rather than 200 ms, to show that the positional stereoblindness also occurs with prolonged viewing. To show that the effect is not specific to judgements of orientation, we used 20 disparity-defined shapes, such as discs, squares or triangles, instead of a clock hand.

Eye position

Eye positions are expressed in Helmholtz angles^{5,8,14}, with positive directions right, up and clockwise (other coordinate systems can mislead in vision studies because they obscure the relation between eye position and retinal correspondence; for example, when eye motion is expressed using quaternions or rotation vectors, converged eyes incycloverge on upgaze^{7,16}, which makes it seem odd that cyclorotated stereograms are seen better). The best way of measuring Helmholtz cyclovergence is the torsional nonius method^{4,9}, but this does not allow the recording of eye position while stereograms are viewed. In six subjects, therefore, we also measured the 3D positions of both eyes using scleral search coils^{25,26}.

Simulations

To compute the white search zones in Fig. 5, we let the eyes move in a pattern known as L2 (ref. 5), in which $C_R = -H_R V/2 - DV/4 = C_L = -H_L V/2 + DV/4$, C and H being the cyclorotation and horizontal angles of the right (R) and left (L) eyes, V their common vertical angle and $D = H_L - H_R$ the vergence angle. Fig. 1 uses a motion pattern intermediate between L2 and Listing's law^{5,7} with $C_R = -H_R V/2 - 0.15DV$, $C_L = -H_L V/2 + 0.15DV$. In both figures the horizontal sizes of the search zones, and to a much lesser extent their heights, depend on the thickness of the stereo-matchable region around the 2D horopter^{2,5}; we assume a thickness of $\pm 3\%$ of the distance from the horopter to the eye because this value, in agreement with experimental data on quantitative stereopsis⁴, implies that subjects should detect horizontal disparities of up to about 1° at the fovea.

We also did simulations using much thicker stereo-matchable regions, up to $\pm 50\%$ of the distance to the eye (with a matchable region this thick, the system would be able to detect disparities of up to 32° at the fovea). The search zones were then much larger, and the per cent reductions achieved by L2 compared to Listing's law were still considerable (for example, about 45% reduction for $\pm 3\%$ thickness, 40% for $\pm 20\%$, 20% for $\pm 50\%$). In Figs 1 and 5, the simulations use 'geometric' horopters; that is, they omit the Hering-Hillebrand deviation and the 1° declinations of corresponding near-vertical meridians^{2,5,15}. Incorporating these small adjustments makes no significant difference to the simulations: L2 still shrinks the search zones by about the same amount.

Received 19 December 2000; accepted 29 January 2001.

- Julesz, B. Binocular depth perception of computer-generated patterns. *Bell. Syst. Tech. J.* **39**, 1152–1162 (1960).
- Ogle, K. *Researches in Binocular Vision* (Saunders, Philadelphia, 1950).
- Rogers, B. J. & Bradshaw, M. F. Does the visual system use the epipolar constraint for matching binocular images? *Invest. Ophthalmol. Vis. Sci.* **37** (Suppl.), 31–25 (1996).
- Stevenson, S. B. & Schor, C. M. Human stereo matching is not restricted to epipolar lines. *Vision Res.* **37**, 2717–2723 (1997).
- Tweed, D. Visual-motor optimization in binocular control. *Vision Res.* **37**, 1939–1951 (1997).
- Gårding, J., Porrill, J., Mayhew, J. E. W. & Frisby, J. P. Stereopsis, vertical disparity and relief transformations. *Vision Res.* **35**, 703–722 (1995).
- Mok, D., Ro, A., Cadera, W., Crawford, J. D. & Vilis, T. Rotation of Listing's plane during vergence. *Vision Res.* **32**, 2055–2064 (1992).
- Van Rijn, L. J. & Van den Berg, A. V. Binocular eye orientation during fixations: Listing's law extended to include eye vergence. *Vision Res.* **33**, 691–708 (1993).
- Howard, I. P., Ohmi, M. & Sun, L. Cyclovergence: a comparison of objective and psychophysical measurements. *Exp. Brain Res.* **97**, 349–355 (1993).
- Hooge, I. Th. C. & van den Berg, A. V. Visually evoked cyclovergence and extended Listing's law. *J. Neurophysiol.* **83**, 2757–2775 (2000).
- Crone, R. A. & Everhard-Halm, Y. Optically induced eye torsion. *Albrecht v. Graefes Arch. klin. exp. Ophthalm.* **195**, 231–239 (1975).
- Hooten, K., Myers, E., Worrall, R. & Stark, L. Cyclovergence: The motor response to cyclodisparity. *Albrecht v. Graefes Arch. klin. exp. Ophthalm.* **210**, 65–68 (1979).

13. Kertesz, A. E., Sullivan, M. J. The effect of stimulus size on human cyclofusional response. *Vision Res.* **18**, 567–571 (1978).
14. Carpenter, R. H. S. *The Movements of the Eyes* (Pion, London, 1988).
15. Helmholtz, H. von. *Handbuch der Physiologischen Optik* (Voss, Hamburg, 1867).
16. Minken, A. W. H. & Van Gisbergen, J. A. M. A three-dimensional analysis of vergence movements at various levels of elevation. *Exp. Brain Res.* **101**, 331–345 (1994).
17. Allen, M. J. The dependence of cyclophoria on convergence, elevation and the system of axes. *Am. J. Optom.* **31**, 297–307 (1954).
18. Kapoula, Z., Bernotas, M. & Haslwanter, T. Listing's plane rotation with convergence: role of disparity, accommodation, and depth perception. *Exp. Brain Res.* **126**, 175–186 (1999).
19. Steffen, H., Walker, M. F. & Zee, D. S. Rotation of Listing's plane with convergence: independence from eye position. *Invest. Ophthalmol. Vis. Sci.* **41**, 715–721 (2000).
20. Hepp, K. On Listing's law. *Commun. Math. Phys.* **132**, 285–292 (1995).
21. Hepp, K. Theoretical explanations of Listing's law and their implication for binocular vision. *Vision Res.* **35**, 3237–3241 (1995).
22. Farell, B. Two-dimensional matches from one-dimensional stimulus components in human stereopsis. *Nature* **395**, 689–693 (1998).
23. Nielsen, K. R. & Poggio, T. Vertical image registration in stereopsis. *Vision Res.* **24**, 1133–1140 (1984).
24. Prazdny, K. Vertical disparity tolerance in random-dot stereograms. *Bull. Psychonom. Soc.* **23**, 413–414 (1985).
25. Robinson, D. A. A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans. Biomed. Eng.* **10**, 137–145 (1963).
26. Tweed, D., Cadera, W. & Vilis, T. Computing three-dimensional eye position quaternions and eye velocity from search coil signals. *Vision Res.* **30**, 97–110 (1990).

Acknowledgements

We thank D. Henriques, I. Howard, E. Klier, P. Nguyen, J. Sharpe, H. Wang, A. Wong and J. Zacher for help with experiments; and K. Beykirch, P. Hallett, C. Hawkins, D. Henriques, H. Misslisch, M. Niemeier and T. Vilis for comments on the manuscript. This study was supported by the Canadian Institutes for Health Research and the Deutsche Forschungsgemeinschaft. D.C. is a CIHR Scholar and D.T. a CIHR Scientist.

Correspondence and requests for materials should be addressed to D.T. (e-mail: douglas.tweed@utoronto.ca).

Leptin-regulated endocannabinoids are involved in maintaining food intake

Vincenzo Di Marzo^{*}, Sravan K. Goparaju[†], Lei Wang^{†‡}, Jie Liu^{†‡}, Sándor Bátkai^{†‡}, Zoltán Járai[†], Filomena Fezza^{*}, Grant I. Miura[§], Richard D. Palmiter[§], Takayuki Sugiyama^{||} & George Kunos^{†‡}

^{*} Endocannabinoid Research Group, Istituto per la Chimica di Molecole di Interesse Biologico, CNR, 80072, Arco Felice, Naples, Italy

[†] Department of Pharmacology and Toxicology, Medical College of Virginia, Virginia Commonwealth University, Richmond, Virginia 23298, USA

[§] Howard Hughes Medical Institute and Department of Biochemistry, University of Washington, Seattle, Washington 98195, USA

^{||} Faculty of Pharmaceutical Sciences, Teikyo University, Sagami, Kanagawa, 199-0195, Japan

Leptin is the primary signal through which the hypothalamus senses nutritional state and modulates food intake and energy balance¹. Leptin reduces food intake by upregulating anorexigenic (appetite-reducing) neuropeptides, such as α -melanocyte-stimulating hormone^{2,3}, and downregulating orexigenic (appetite-stimulating) factors, primarily neuropeptide Y⁴. Genetic defects in anorexigenic signalling, such as mutations in the melanocortin-4 (ref. 5) or leptin receptors⁶, cause obesity. However, alternative orexigenic pathways maintain food intake in mice deficient in neuropeptide Y⁷. CB1 cannabinoid receptors⁸ and the endocannabinoids anandamide and 2-arachidonoyl glycerol are present in the hypothalamus⁹, and marijuana¹⁰ and anandamide^{11,12} stimulate food intake. Here we show that following temporary food restriction, CB1 receptor knockout mice eat

less than their wild-type littermates, and the CB1 antagonist SR141716A reduces food intake in wild-type but not knockout mice. Furthermore, defective leptin signalling is associated with elevated hypothalamic, but not cerebellar, levels of endocannabinoids in obese *db/db* and *ob/ob* mice and Zucker rats. Acute leptin treatment of normal rats and *ob/ob* mice reduces anandamide and 2-arachidonoyl glycerol in the hypothalamus. These findings indicate that endocannabinoids in the hypothalamus may tonically activate CB1 receptors to maintain food intake and form part of the neural circuitry regulated by leptin.

CB1^{-/-} and CB1^{+/+} mice were made as described previously^{13,14} and were maintained on a reverse light/dark cycle for measurement of food intake. After 18 h of fasting, the animals received a single intraperitoneal injection of vehicle or drug 10 min before the beginning of the dark cycle, and food intake was measured for the indicated period of time. After vehicle treatment, food intake was significantly lower in CB1^{-/-} than in CB1^{+/+} mice. When CB1^{+/+} mice were treated with 3 μ g g⁻¹ of the selective CB1 receptor antagonist SR141716A (ref. 15), food intake was significantly reduced to the same level as in vehicle-treated CB1^{-/-} mice, whereas in CB1^{-/-} mice SR141716A did not affect food intake (Fig. 1). These results indicate that endogenous cannabinoids acting at CB1 receptors may be involved in maintaining food intake in mice made hyperphagic by brief food deprivation.

As leptin is known to downregulate the expression in the hypothalamus of orexigenic peptides such as neuropeptide Y (NPY)⁴, orexins¹⁶ and melanin concentrating hormone¹⁷, we investigated whether it may similarly regulate hypothalamic endocannabinoids. A single intravenous injection of 125 or 250 μ g of recombinant mouse leptin into normal Sprague–Dawley rats resulted in around 40–50% reductions in the hypothalamic levels of both anandamide and 2-arachidonoyl glycerol (2-AG) within 30 min, compared with levels in vehicle-treated controls (Fig. 2). By contrast, defective leptin signalling in obese Zucker rats was associated with elevated 2-AG levels in the hypothalamus compared with non-obese controls. The hypothalamic levels of anandamide and palmitoyl ethanolamide (PEA) were not significantly different in Zucker rats and their controls (Fig. 3a). In young (6–8-week-old) obese *db/db* mice with defective leptin receptors, hypothalamic levels of both 2-AG and anandamide were higher than in their

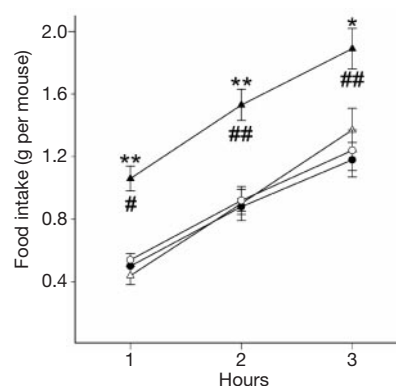


Figure 1 Food intake in CB1^{-/-} mice and their CB1^{+/+} littermates in the absence and presence of the CB1 receptor antagonist SR141716A. Cumulative food intake was measured in 18-h food-restricted CB1^{-/-} (circles) and CB1^{+/+} mice (triangles). The animals received a single intraperitoneal injection of vehicle (filled symbols) or 3 μ g g⁻¹ SR141716A (open symbols) 10 min before the start of the testing period. Means $\pm$ s.e., $n = 9$ CB1^{+/+} and 5 CB1^{-/-} mice. Significance of difference from corresponding values in the presence of SR141716A in the same group of animals (* $P < 0.05$, ** $P < 0.01$), or from corresponding values in vehicle-treated CB1^{-/-} mice (# $P < 0.05$, ## $P < 0.01$) calculated by analysis of variance followed by Tukey's test. Food intake in non-fasted animals was much lower; it was similar in CB1^{+/+} (0.19 $\pm$ 0.05, 0.33 $\pm$ 0.06, 0.46 $\pm$ 0.06 g per mouse, $n = 8$) and CB1^{-/-} mice (0.16 $\pm$ 0.02, 0.33 $\pm$ 0.04, 0.45 $\pm$ 0.05 g per mouse, $n = 8$), at 1, 2 and 3 h, respectively.

[‡] Present address: National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health, MSC-8115, Bethesda, Maryland 20892-8115, USA.

C57BL/6J controls, whereas the levels of PEA were similar in the two groups (Fig. 3b). In young *ob/ob* mice, which lack leptin, hypothalamic levels of 2-AG, but not anandamide, were higher than in their C57BLKS/J controls (Fig. 3c). In *ob/ob* mice treated with leptin ($2 \mu\text{g g}^{-1}$ intravenously) hypothalamic 2-AG levels were similar to levels measured in lean controls, and anandamide declined to undetectable ($<1.0 \text{ pmol g}^{-1}$) levels. These findings strongly suggest negative regulation of endocannabinoids by leptin, which appears to be specific for endocannabinoids and the hypothalamus, as the hypothalamic levels of PEA were similar and there was no difference in the cerebellar levels of anandamide (18.2 ± 1.0 versus $16.1 \pm 0.5 \text{ pmol g}^{-1}$, $P > 0.2$) or 2-AG (7.5 ± 0.3 versus $8.1 \pm 0.6 \text{ nmol g}^{-1}$, $P > 0.4$) between *db/db* mice and their controls, respectively.

In the absence of a neuronal storage mechanism for endocannabinoids, the effect of leptin on endocannabinoid levels in the hypothalamus may reflect decreased synthesis, increased degradation or both. Hypothalamic activity of the enzyme fatty acid amide hydrolase (FAAH), which hydrolyses both anandamide and 2-AG¹⁸, was unaffected by leptin (1.76 ± 0.10 versus $1.73 \pm 0.04 \text{ nmol per min per mg protein}$ in preparations from three control and three leptin-treated rats; $P > 0.5$). On the other hand, the hypothalamic levels of the direct biosynthetic precursor of anandamide, N-Ar-PE (ref. 19), increased significantly from 81 ± 8 to $221 \pm 8 \text{ pmol g}^{-1}$ ($n = 3$, $P < 0.001$) following treatment with leptin.

This indicates that leptin may downregulate the activity of the phospholipase D (PLD) enzyme that catalyses the conversion of N-Ar-PE into anandamide^{19,20}, leading to its accumulation. However, the activity of the N-Ar-PE-specific PLD in the hypothalamus was not significantly altered by treatment with $250 \mu\text{g}$ leptin (11.2 ± 2.2 versus $14.9 \pm 3.1 \text{ pmol per min per mg protein}$, $n = 6$, $P > 0.3$). Leptin may trigger or inhibit the release of putative soluble inhibitors or activators, respectively, of the N-Ar-PE-specific PLD, and such soluble substances may be lost in the *in vitro* assay system used to measure enzyme activity. For example, leptin acutely decreases low-glucose-induced mobilization of intracellular calcium in the arcuate nucleus²¹, and may thereby reduce the activity of the N-Ar-PE-specific PLD in a way that could not be reproduced in a cell-free system *in vitro*.

We also investigated the effect of leptin on the biosynthesis of 2-AG. We assayed the activity of diacyl glycerol (DAG) lipase, the enzyme responsible for the generation of 2-AG from DAG, in hypothalamic homogenates by the conversion of 1-stearoyl-2-[^{14}C]arachidonyl-*sn*-glycerol to [^{14}C]2-AG in the absence and presence of a DAG lipase inhibitor²². Like that of PLD, DAG lipase activity was similar in preparations from rats pretreated with a

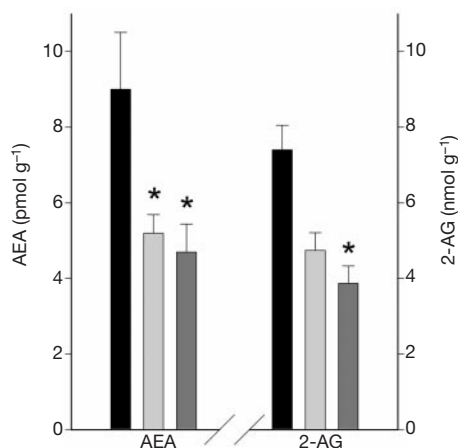


Figure 2 Hypothalamic levels of anandamide and 2-arachidonoyl glycerol (2-AG). Black, control rats; light grey, rats treated with $125 \mu\text{g}$ leptin; dark grey, rats treated with $250 \mu\text{g}$ leptin (means $\pm$ s.e. from six separate experiments; asterisk, $P < 0.05$).

single intravenous injection of saline or $250 \mu\text{g}$ leptin (3.12 ± 0.22 versus $4.27 \pm 0.48 \text{ pmol per min per mg protein}$, $n = 3$, $P > 0.1$). We also measured the hypothalamic levels of *sn*-2-arachidonate-containing DAGs, the immediate precursors of 2-AG. They were markedly reduced by leptin treatment, from 67.5 ± 10.6 to $19.6 \pm 3.3 \text{ nmol per g wet weight}$. This indicates that, unlike the situation with anandamide, the target of leptin may be upstream from the generation of DAG, possibly at the level of a phosphatidylinositol-specific PLC, which are calcium sensitive. The exact mechanisms by which leptin reduces endocannabinoid levels in the hypothalamus therefore remain to be clarified.

The ability of SR141716A to reduce food intake in rodents has been reported^{23,24}. However, the well-documented inverse agonist activity of SR141716A (ref. 25) and its possible nonspecific effects²⁶ make it difficult to conclude from those results that food intake is tonically stimulated by endocannabinoids. Unambiguous evidence for such a conclusion is provided by our findings that food intake is significantly lower in CB1 receptor-deficient mice than in their wild-type controls, and that SR141716A can reduce food intake in wild-type but not in CB1 receptor-deficient mice. The finding that the much lower food intake in animals with unlimited access to food is similar in CB1^{+/+} and CB1^{-/-} mice (Fig. 1) is compatible with the lack of difference in body weight between the two groups¹³ and suggests that, in the absence of food deprivation, other orexigenic signals may compensate for the loss of endocannabinoid signalling.

Genetic defects in anorexigenic signalling pathways, such as the absence of leptin or melanocortin-4 receptors, lead to obesity^{5,6}. However, the lack of hypophagia and of a lean phenotype in NPY-deficient animals⁷ indicates that there is more redundancy in

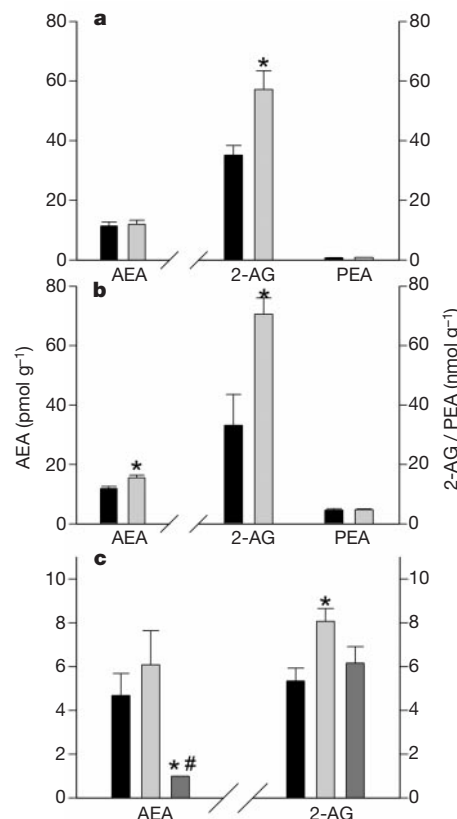


Figure 3 Hypothalamic levels of anandamide (AEA), 2-arachidonoyl glycerol (2-AG) and palmitoyl ethanolamide (PEA). **a**, Obese Zucker (*fa/fa*) rats (grey, $n = 12$) and lean controls (black, $n = 9$); **b**, obese *db/db* mice (grey, $n = 4$) and lean controls (black, $n = 4$); **c**, obese *ob/ob* mice treated with vehicle (light grey, $n = 8$) or $2 \mu\text{g g}^{-1}$ leptin intravenously (dark grey, $n = 5$), and lean controls (black, $n = 8$). Means $\pm$ s.e.; significance of difference from the corresponding values in lean controls (* $P < 0.05$) or from vehicle-treated *ob/ob* mice (# $P < 0.05$) is shown.

orexigenic signalling. We tested whether endocannabinoids may be the main factors compensating for the lack of NPY by comparing the ability of SR141716A to inhibit food intake in food-restricted NPY-deficient mice and their wild-type littermates. As illustrated in Fig. 4, a single intraperitoneal injection of $3 \mu\text{g g}^{-1}$ SR141716A 10 min before the dark cycle caused a marked reduction in food intake, which was similar in the two groups. This may indicate that endocannabinoids are involved in maintaining food intake in both the absence and the presence of NPY, and that the two systems operate essentially independently of one another. Simultaneous blockade of more than one orexigenic pathway may be necessary to achieve clinically relevant, long-term reductions in food intake in obese individuals.

The increased hypothalamic levels of endocannabinoids in genetically obese animals may indicate that these substances contribute to the hyperphagia that leads to obesity. We measured the effect of SR141716A on food intake in 5–6-week-old obese *db/db*

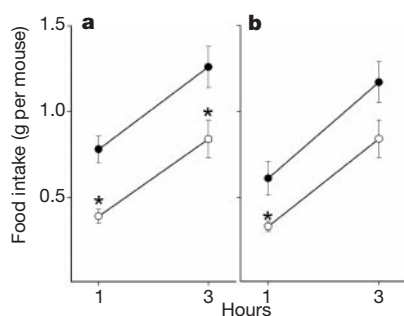


Figure 4 Effect of SR141716A on food intake in NPY-deficient mice. Food intake in NPY-deficient mice (a) and their controls (b) after acute treatment with vehicle (filled circles) or $3 \mu\text{g g}^{-1}$ SR141716A intraperitoneally (open circles). Mice kept under normal light/dark cycle were food deprived for 24 h before testing during the beginning of the dark phase. Means $\pm$ s.e. from eight animals in each treatment group. Asterisk, significant difference ($P < 0.05$) from vehicle-treated group.

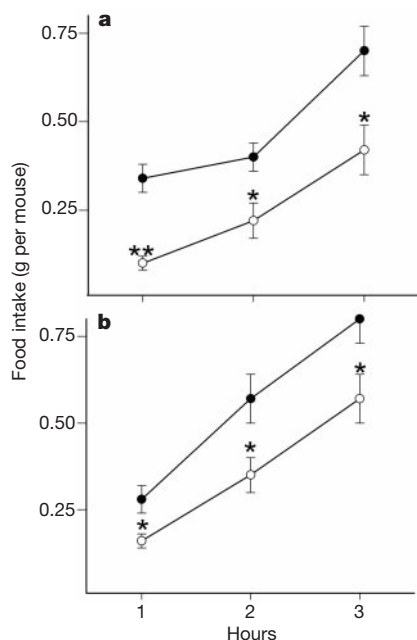


Figure 5 Effect of SR141716A on food intake in obese mice. Food intake in non-fasted *db/db* (a) and *ob/ob* mice (b) after acute treatment with vehicle (filled circles) or $3 \mu\text{g g}^{-1}$ SR141716A intraperitoneally 10 min before the test period. Means $\pm$ s.e. from 6–12 animals in each group. Asterisks indicate significant difference (* $P < 0.05$, ** $P < 0.01$) from the vehicle-treated group.

and *ob/ob* mice with unrestricted access to food. Intraperitoneal injection of $3 \mu\text{g g}^{-1}$ SR141716A caused a significant, acute reduction in food intake in both types of obese mice (Fig. 5). Furthermore, daily treatment of four *db/db* mice with the same dose of SR141716A for seven days resulted in a reduction in body weight compared with four vehicle-treated *db/db* mice. On day four of the treatment, body weight in the two groups was $98.1 \pm 1.1\%$ versus $102.0 \pm 0.8\%$ of day 0 values, respectively ($P < 0.05$), but by day 6 catch-up growth in the SR141716A-treated animals had eliminated the difference. This may indicate that endocannabinoids, with other leptin-regulated orexigenic signals, contribute to overeating in the development of obesity.

Hypothalamic endocannabinoids appear to be under negative control by leptin and may be considered to belong to the growing family of orexigenic mediators, including NPY⁷, the orexins^{16,27} and melanin concentrating hormone^{17,28}. However, the cell groups in which endocannabinoids are synthesized and their neuronal connections with other hypothalamic nuclei that form the circuitry controlling food intake and satiety remain to be clarified. □

Methods

Animals

We bred CB1 receptor-deficient (*CB1*^{-/-}) mice and their homozygote littermates (*CB1*^{+/+} mice) as described^{13,14}. Details about the NPY-deficient mice and their controls have been reported⁷. Obese *db/db* and *ob/ob* mice with their respective C57BL/6J and C57BLKS/J controls and obese Zucker (*fa/fa*) rats with their controls were obtained from Jackson Laboratories. Sprague–Dawley rats were from Harlan.

Food intake measurement

We maintained male and female mice on a reverse light/dark cycle (lights on 21:00 to 09:00, lights off 09:00 to 21:00), with one animal per cage. The animals were fasted for 18 h before testing, except when noted otherwise. NPY-deficient mice and their controls were tested during their normal dark cycle following a 24-h fast. Ten minutes before the start of the test period, the animals received an intraperitoneal injection of vehicle or $3 \mu\text{g g}^{-1}$ SR141716A (*N*-(piperidin-1-yl)-5-(4-chlorophenyl)-1-(2,4-dichlorophenyl)-4-methyl-1H-pyrazole-3-carboxamide HCl). Ten minutes later a food tray was placed in their cage, and cumulative food intake was measured by weighing the residual food in the tray at the indicated intervals and correcting for spillage, which was minimal.

Measurement of endocannabinoids in the brain

Animals were anaesthetized with ether and decapitated, and the brain was removed. We dissected the hypothalami by parallel coronal cuts rostrally at the edge of the optical chiasm and caudally through the mamillary bodies, then sagittal cuts 2 mm lateral from the midline (1.5 mm in mice), and finally a horizontal cut at the top of the third ventricle. The cerebellum was also removed and both tissues were immediately frozen in liquid nitrogen, within 1 min of decapitation. The hypothalami from two rats or four mice were pooled and extracted with 3 vol of chloroform/methanol 2:1 (v/v) containing 1 nmol *d*₈-anandamide (Cayman Chemicals), 2 nmol of *d*₈-2-AG (Deva Biotech) and 10 nmol of *N*-palmitoyl-phosphatidylethanolamine (Sigma) as external standards. The extracts were fractionated by chromatography on silica gel mini-columns eluted with chloroform/methanol 9:1 (v/v), as described⁹. We analysed pre-purified lipids by high-performance liquid chromatography–atmospheric pressure chemical ionization mass spectrometry (HPLC–APCI–MS). HPLC separation was through a Supelcosil LC-18 column (150 mm $\times$ 4.6 mm, 5 μm particle size), using a mobile phase of methanol/water/acetic acid 85:15:0.2 (v/v/v) at a flow rate of 1 ml min⁻¹. The MS was equipped with a Z-Spray APCI source operating in the (+) APCI mode (source temperature 120 °C, probe temperature 110 °C). N₂ was used as both the drying and nebulizing gas, with flow and probe position adjusted daily for optimum sensitivity. Both the column and the samples were maintained at 25 °C. Anandamide and 2-AG in their protonated (*M*+1) form were identified as peaks with the appropriate *m/z* values, and quantified by the isotope dilution method, whereas palmitoyl ethanolamide was quantified by comparison with its external synthetic standard run under the same conditions. Values are expressed as pmol or nmol per g wet tissue.

We purified *N*-Ar-PE as described⁹ and quantified it by the amount of anandamide released through its digestion with *Streptomyces chromofuscus* PLD, which cleaves quantitatively all *N*-acyl-phosphatidylethanolamines^{9,18}. We quantified the anandamide released by HPLC–APCI–MS, as described above. The consistency of the yield of *N*-Ar-PE purification and hydrolysis was verified by parallel LC–MS measurement of palmitoyl-ethanolamide released from *N*-palmitoyl-phosphatidylethanolamine, which was added to each sample as an external standard before the purification. Values are expressed as pmol per g wet tissue.

Enzyme assays

We measured FAAH activity as described²⁹. Rats were killed by decapitation 30 min after the intravenous injection of vehicle or 250 μg leptin. Membrane fractions from the

hypothalamus were incubated with 100 μM [^{14}C]anandamide in Tris-HCl, pH 9, at 37 °C for 15 min, after which 2 vol of chloroform/methanol 2:1 was added to the incubation mixture. The [^{14}C]ethanolamine produced from the enzymatic reaction was quantified by counting the aqueous phase, and was used as a measure of FAAH activity.

The activity of the N-Ar-PE specific PLD was measured by using [^3H]N-Ar-PE (5 mCi mmol^{-1}) as substrate²⁰. We prepared whole homogenates from the hypothalamus of rats killed within 30 min of an intravenous injection of vehicle or 250 μg leptin. The homogenates were incubated with [^3H]N-Ar-PE (100,000 c.p.m., 20 μM) in 1 ml of 20 mM HEPES-NaOH buffer, pH 7.4, containing 2 mM dithiothreitol, 2 mM CaCl_2 , 2 mM Triton-X100 and 1 mM phenylmethyl sulphonyl fluoride for 30 min at 37 °C, and the reaction terminated by the addition of 2 vol of chloroform/methanol 2:1. The organic phase was fractionated by thin layer chromatography, developed with chloroform/methanol/ NH_4OH (80:20:2, v/v/v), and the band corresponding to the R_f (migration index) value of synthetic anandamide was scraped off the plate and quantified by liquid scintillation spectrometry.

Diacyl glycerol lipase activity was measured in rat hypothalamic homogenates as described²². We quantified *sn*-2-arachidonate-containing DAGs by the arachidonoyl methyl ester (AME) produced from their *trans*-esterification with methanol. We purified hypothalamic lipid extracts by thin layer chromatography on silica plates eluted with diethyl ether/petroleum ether/ammonia (50:50:1, by vol), under conditions allowing the separation of DAGs ($R_f = 0.5$) from triacylglycerols ($R_f = 1$), phospholipids ($R_f = 0.1$) and free fatty acids ($R_f = 0.25$). Bands corresponding to a standard of 1-oleoyl-2-arachidonoyl-glycerol were eluted from silica gel, lyophilized and reacted with 0.5 ml anhydrous methanol in the presence of 3 mg anhydrous sodium carbonate overnight at room temperature. The fatty acid methyl esters produced from the reaction (> 90% yield) were purified by phase extraction, mixed with 1 μg [$^2\text{H}_6$]AME and analysed by gas chromatography-electron impact mass spectrometry (GC-EIMS) as described⁹. Selected fragments were at $m/z = 326$ and 318, corresponding to the molecular ions of [$^2\text{H}_6$]AME and AME, and at $m/z = 255$ and 247, corresponding to the loss of the terminal *n*-pentyl chain. AME was quantified by isotope dilution with [$^2\text{H}_6$]AME, using the GC signal area ratios of the peaks with $m/z = 255$ and 247.

Received 27 November 2000; accepted 15 January 2001.

- Friedman, J. M. & Halaas, J. L. Leptin and the regulation of body weight in mammals. *Nature* **395**, 763–770 (1998).
- Schwartz, M. W. *et al.* Leptin increases hypothalamic pro-opiomelanocortin mRNA expression in the rostral arcuate nucleus. *Diabetes* **46**, 2119–2123 (1997).
- Seeley, R. J. *et al.* Melanocortin receptors in leptin effects. *Nature* **390**, 349–351 (1997).
- Stephens, T. W. *et al.* The role of neuropeptide Y in the antiobesity action of the obese gene product. *Nature* **377**, 530–534 (1995).
- Huszar, D. *et al.* Targeted disruption of the melanocortin-4 receptor results in obesity in mice. *Cell* **88**, 131–141 (1997).
- Chen, H. *et al.* Evidence that the diabetes gene encodes the leptin receptor: identification of a mutation in the leptin receptor gene in *db/db* mice. *Cell* **84**, 491–495 (1996).
- Erickson, J. C., Clegg, K. E. & Palmiter, R. D. Sensitivity to leptin and susceptibility to seizures in mice lacking neuropeptide Y. *Nature* **381**, 415–421 (1996).
- Herkenham, M. *et al.* Characterization and localization of cannabinoid receptors in rats: a quantitative *in vitro* autoradiographic study. *J. Neurosci.* **11**, 563–583 (1991).
- Gonzalez, S. *et al.* Identification of endocannabinoids and cannabinoid CB1 receptor mRNA in the pituitary gland. *Neuroendocrinology* **70**, 137–145 (1999).
- Mechoulam, R., Hanus, L. & Fride, E. Towards cannabinoid drugs—revisited. *Prog. Med. Chem.* **35**, 199–243 (1998).
- Williams, C. M. & Kirkham, T. C. Anandamide induces overeating: mediation by central cannabinoid (CB1) receptors. *Psychopharmacology* **143**, 315–317 (1999).
- Hao, S., Avraham, Y., Mechoulam, R. & Berry, E. M. Low dose anandamide affects food intake, cognitive function, neurotransmitter and corticosterone levels in diet-restricted mice. *Eur. J. Pharmacol.* **392**, 147–156 (2000).
- Zimmer, A., Zimmer, A. M., Hohmann, A. G., Herkenham, M. & Bonner, T. I. Increased mortality, hypoactivity, and hypoalgesia in cannabinoid CB1 receptor knockout mice. *Proc. Natl Acad. Sci. USA* **96**, 5780–5785 (1999).
- Járai, Z. *et al.* Cannabinoid-induced mesenteric vasodilation through an endothelial site distinct from CB1 or CB2 receptors. *Proc. Natl Acad. Sci. USA* **96**, 14136–14141 (1999).
- Rinaldi-Carmona, M. *et al.* SR141716A, a potent and selective antagonist of the brain cannabinoid receptor. *FEBS Lett.* **350**, 240–244 (1994).
- Lopez, M. *et al.* Leptin regulation of prepro-orexin and orexin receptor mRNA levels in the hypothalamus. *Biochem. Biophys. Res. Commun.* **269**, 41–45 (2000).
- Sahu, A. Evidence suggesting that galanin (GAL), melanin-concentrating hormone (MCH), neurotensin (NT), proopiomelanocortin (POMC) and neuropeptide Y (NPY) are targets for leptin signaling in the hypothalamus. *Endocrinology* **139**, 795–798 (1998).
- Di Marzo, V., Melck, D., Bisogno, T. & De Petrocellis, L. Endocannabinoids: endogenous cannabinoid receptor ligands with neuromodulatory action. *Trends Neurosci.* **21**, 521–528 (1998).
- Di Marzo, V. *et al.* Formation and inactivation of endogenous cannabinoid anandamide in central neurons. *Nature* **372**, 686–691 (1994).
- Sugiura, T. *et al.* Transacylase-mediated and phosphodiesterase-mediated synthesis of N-arachidonoyl-ethanolamine, an endogenous cannabinoid-receptor ligand, in rat brain microsomes. Comparison with synthesis from free arachidonic acid and ethanolamine. *Eur. J. Biochem.* **240**, 53–62 (1996).
- Funahashi, H. *et al.* The effect of leptin on feeding-regulating neurons in the rat hypothalamus. *Neurosci. Lett.* **264**, 117–120 (1999).
- Bisogno, T. *et al.* Phosphatidic acid as the biosynthetic precursor of the endocannabinoid 2-arachidonoylglycerol in intact mouse neuroblastoma cells stimulated with ionomycin. *J. Neurochem.* **72**, 2113–2119 (1999).
- Colombo, G. *et al.* Appetite suppression and weight loss after the cannabinoid antagonist SR 141716A. *Life Sci.* **63**, PL113–PL117 (1998).
- Chaperon, F., Soubrie, P., Puech, A. J. & Thiebot, M.-H. Involvement of central cannabinoid (CB1)

- receptors in the establishment of place conditioning in rats. *Psychopharmacology* **135**, 324–332 (1998).
- Bouaboula, M. *et al.* A selective inverse agonist for central cannabinoid receptor inhibits mitogen-activated protein kinase activation stimulated by insulin or insulin-like growth factor. 1. Evidence for a new model of receptor/ligand interactions. *J. Biol. Chem.* **272**, 22330–22339 (1997).
- White, R. & Hiley, C. R. The actions of the cannabinoid receptor antagonist, SR141716A, in rat isolated mesenteric artery. *Br. J. Pharmacol.* **125**, 689–696 (1998).
- Sakurai, T. *et al.* Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
- Shimada, M., Tritos, N. A., Lowell, B. B., Flier, J. S. & Maratos-Flier, E. Mice lacking melanin-concentrating hormone are hypophagic and lean. *Nature* **396**, 670–674 (1998).
- Di Marzo, V. *et al.* Biosynthesis and inactivation of the endocannabinoid 2-arachidonoylglycerol in circulating and tumoral macrophages. *Eur. J. Biochem.* **264**, 258–267 (1999).

Acknowledgements

We thank A. Zimmer for heterozygote breeding pairs for the CB1^{-/-} and CB1^{+/-} mice. This study was supported by grants from the NIH to G.K. Z.J. and S.B. were supported by fellowships from Sanofi Research (Montpellier, France) and the Martin Rodbell visiting program, respectively. V.D.M. was the recipient of a Human Frontier Science Program short-term fellowship.

Correspondence and requests for materials should be addressed to G.K. (e-mail: gkunos@mail.nih.gov).

Induction of the mammalian node requires Arkadia function in the extraembryonic lineages

Vasso Episkopou, Ruth Arkell^{*†}, Paula M. Timmons, James J. Walsh^{*}, Rebecca L. Andrew & Daniel Swan^{*†}

Mammalian Neurogenesis, MRC Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, Du Cane Road, London W12 0NN, UK

[†] These authors contributed equally to this work.

The early mammalian embryo is patterned by signals emanating from extraembryonic and embryonic signalling centres, most notably the anterior visceral endoderm (AVE) and the node, respectively¹. The AVE is responsible for anterior development, whereas further axis specification depends on the node, the equivalent of Spemann's organizer^{2,3}. Formation of the node, at the anterior primitive streak, depends on expression of the transcription factor HNF3 β (ref. 4). However, both the source and the nature of the signals responsible for inducing the node have been unknown. Here we describe a recessive lethal mutation, *arkadia*, generated using gene-trap mutagenesis. Mutant embryos establish an AVE but fail to maintain anterior embryonic structures and lack a node. The mutation has disrupted the *Arkadia* gene, which encodes a putative intracellular protein containing a RING domain. *Arkadia* is essential for HNF3 β expression in the anterior primitive streak. Analysis with chimaeras, however, shows that *Arkadia* functions within extraembryonic tissues, revealing that these are required to induce the node. Furthermore, our experiments show that *Arkadia* interacts genetically with the transforming growth factor (TGF) β -like factor Nodal^{5–7}, implying that Nodal mediates the function of *Arkadia* in node induction.

We generated recessive lethal gene-trap insertion mutations in genes expressed in embryonic stem (ES) cells⁸ using the retroviral ROSA β geo gene-trap vector (GT). We describe one such mutation, *arkadia*. The promoterless GT-vector marker, β geo, is expressed widely, indicating that the promoter of the endogenous *Arkadia* gene is ubiquitously active (Fig. 1a). The homozygous *arkadia*

^{*}Present addresses: Mammalian Genetics Unit, MRC, Harwell, OX11 0RD, UK (R.A., J.W.); Centre for Genome Research, The University of Edinburgh, Edinburgh, EH9 3JQ, UK (D.S.)

mutant embryos (*arkadia* embryos) exhibit specific abnormalities, apparent from early somite stages. We studied the phenotype in an inbred 129Sv/Ev genetic background and found that *arkadia* embryos rarely develop beyond 15 somites, have a reduced head, fail to undergo turning and die at midgestation (Fig. 1a).

We first examined the extent of neurectoderm truncation and anterior-posterior patterning in *arkadia* embryos between the headfold and 15-somite stages (Fig. 1) using the expression of the following neurectoderm marker genes: *Sox3*, an early neural plate marker⁹; *Otx2* and *Pax2*, markers for forebrain and midbrain^{10,11}; and *Follistatin*, a hindbrain marker¹². Presomite *arkadia* embryos develop normal headfolds, as shown by *Sox3* expression (Fig. 1b) and histological sectioning (data not shown). At somite stages, though, anterior development arrests in *arkadia* embryos. The headfolds begin to appear narrower than normal, the neural groove is indistinct and, beginning with the most anterior structures, embryos progressively lose all tissues rostral to the hindbrain. This was confirmed by our finding of a progressive diminution of the forebrain and midbrain domains of expression of both *Otx2* (Fig. 1d–f) and *Pax2* (Fig. 1g–i). The most posterior marker, *Follistatin*, showed that the neurectoderm truncation extends to the anterior limit of its expression at rhombomere 2 (Fig. 1c). In conclusion, presomite *arkadia* embryos seem to initiate normal neurectoderm development but anterior forebrain and midbrain tissues are not maintained.

Direct observation and the somite-specific expression of *Follistatin*¹² (Fig. 1c) and *Cerberus related* (*Cer1*)¹³ (Fig. 2a) show that the first and most anterior somites develop fused across the midline, indicating a possible defect in the development of structures in the anterior midline. This was confirmed by *Shh* expression, which normally occurs in the mesendoderm (notochord/prechordal plate), the floorplate of the neural tube and throughout the gut endoderm¹⁴ (Fig. 2c, d). In 9.5-day-postcoitum (d.p.c.) *arkadia* embryos, *Shh* is present in the midgut and hindgut (Fig. 2c), but is absent from the foregut and mesendoderm. Histological sectioning of the mutant embryos showed that the tissues are absent (Fig. 2e and data not shown). As mesendoderm and endoderm derive from

the node and the primitive streak respectively, we used the specific marker *Brachyury*¹⁵ to show that mutant embryos have a primitive streak that is indistinguishable from that of wild-type embryos, but both node and notochord are absent (Fig. 2b).

Breeding data and Southern blot analysis showed that the *arkadia* mutation is linked to a single GT provirus (data not shown). Analysis of the genomic sequence adjacent to the GT insertion revealed precise retroviral integration. We used 5' rapid amplification of complementary DNA ends (RACE) polymerase chain reaction with reverse transcriptase (RT-PCR) to obtain the endogenous exons spliced onto β geo, and retrieved cDNAs corresponding to endogenous transcripts. The cDNAs contained an open reading frame (ORF) of 989 amino acids encoding a protein, Arkadia (Fig. 3a) (accession number AF330197), which was found to be highly conserved in other vertebrates (Table 1). Database analysis of the predicted amino-acid sequence of Arkadia revealed no similarity with known proteins. However, Arkadia possesses a RING-H2 finger motif at its carboxy-terminal end (Fig. 3a). These are short zinc-binding motifs (40–60 amino acids) found in many proteins with otherwise diverse structures and cellular localization that are thought to be involved in the ubiquitin–proteasome pathway¹⁶. Other features of the predicted protein are two putative nuclear localization signals and low-complexity amino-acid repeats (Fig. 3a). Preliminary experiments, where Arkadia has been fused to green fluorescent protein (GFP) and expressed in COS cells (data not shown) or frog embryos¹⁷, suggest nuclear or intracellular localization.

The two endogenous *Arkadia* exons (Fig. 3d) that are spliced onto β geo correspond to the 5' untranslated region (UTR) of the cDNAs, indicating that the GT insertion occurred within an intron close to the 5' end and to the promoter of the *Arkadia* gene. This was confirmed by a β geo-specific probe that hybridizes to an approximately 4.3-kb transcript (Fig. 3d), which is around 0.3 kb larger than β geo itself. Therefore, this retroviral messenger RNA probably initiates at the endogenous promoter and contains exons located between the promoter and the polyadenylation site at the end of

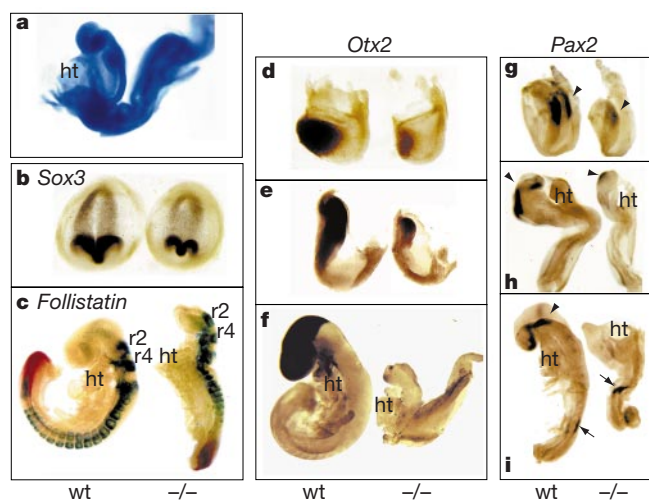


Figure 1 Neurectoderm development in *arkadia* embryos. **a**, X-gal-stained 9.5-d.p.c. *arkadia* embryo showing head truncation and widespread expression of lacZ reporter. **b–i**, WISH of wild-type (wt, left) and *arkadia* (right) embryos with probes as indicated. **a, c–i**, Lateral views with anterior uppermost or to the left. **b**, Dorsal view of headfold (HF) stage with anterior downwards. **c**, Double WISH for *Follistatin* (blue) and *Brachyury* (red) on wild-type and *arkadia* embryos at 9.5 d.p.c. **d–f**, Neural plate (NP), HF and 20-somite stages. **g–i**, HF, 8-somite and 12-somite stages. ht, heart; r, rhombomere; $-/-$, *arkadia* homozygous mutant; wt, wild type (*Arkadia* $+/+$ or $+/-$); arrows, nephrogenic cords; arrowheads, midbrain. In older *arkadia* embryos, there is preferential loss of anterior (forebrain–midbrain **c, i, f**) compared to posterior structures (nephrogenic cords in **i**).

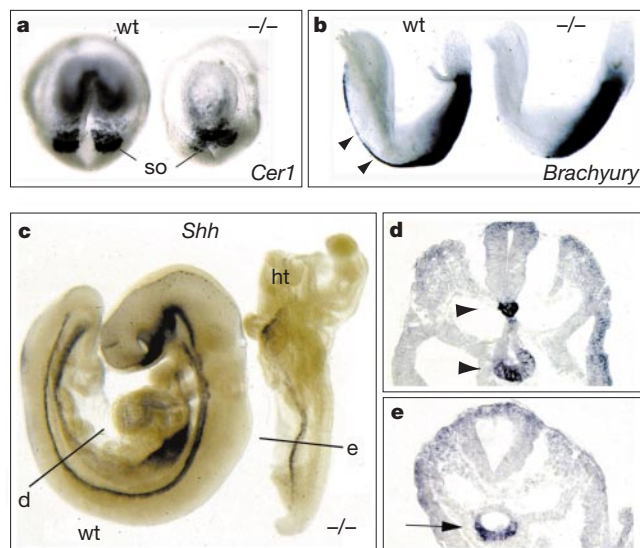


Figure 2 Mesendoderm and definitive endoderm in *arkadia* embryos. Embryos (**a–c**) and sections (**d, e**) of wild-type (wt) and *arkadia* embryos subject to WISH with probes as indicated. **a**, Frontal–ventral view at approximately 3-somite stage. *Cer1* expression spanning the midline shows fusion of anterior somites in the mutant. **b**, Lateral view of HF stage, anterior to the left, showing absence of notochord in the mutant. **c–e** *Shh* expression at 9.5 d.p.c., marking mesendodermal tissues (notochord and prechordal plate), gut endoderm and floorplate in wild-type embryo but only mid- and hind-gut in mutant. ht, heart; so, somite; $-/-$, *arkadia* homozygous mutant; wt, wild type (*Arkadia* $+/+$ or $+/-$); arrow, gut; arrowheads, notochord.

βgeo. Comparison of the *Arkadia* cDNAs with 20 kb of genomic sequence around the provirus revealed that all coding exons lie more than 6.5 kb downstream of the insertion (Fig. 3b). The GT insertion maps to mouse chromosome 9 (Fig. 3c) and the phenotype cannot be correlated with any known mutations in this region. The expression of both the *βgeo* marker gene of the GT vector (Fig. 1a) and the endogenous *Arkadia* gene is the same; both are present in all tissues examined at stages from ES cells to midgestation (Fig. 3d–f and data not shown) and in the adult (data not shown).

Three additional lines of evidence show that transcription through the *Arkadia* ORF has been inactivated by the GT vector insertion. First, whole-mount *in situ* hybridization (WISH) with *Arkadia* coding region probes shows broad expression in wild-type embryos, whereas expression is completely absent in mutants (Fig. 3f). Second, using *Arkadia* coding region probes, abundant 4.5–5.5-kb transcripts are detected in northern blots of RNA from wild-type and heterozygous ES cells. However, no transcripts are detected in RNA from homozygous mutant ES cells (Fig. 3d). Finally, RT–PCR experiments detect transcription through a control *α-galactosidaseA* housekeeping gene (data not shown), and through 5' untranslated region *Arkadia* exons located 5' of the GT insertion in wild-type, heterozygous and mutant embryo RNA. However, RT–PCR with two sets of nested primers spanning the coding region of the *Arkadia* transcript failed to amplify a product in mutant RNA, but succeeded with RNA from wild-type or heterozygous embryos, at stages around the onset of the phenotype (Fig. 3e and data not shown). Therefore, we consider *arkadia* to be a complete loss-of-function mutation.

To investigate the developmental origins of the *arkadia* phenotype we used PCR to genotype the embryos and extended our analysis to earlier stages, when abnormalities are not yet visible. We investigated the presence of node precursor cells in the mutant

Table 1 Comparison of vertebrate *Arkadia* amino-acid sequences

	Arkadia	Hs Arkadia	cArkadia	XArkadia
Arkadia	–			
Hs Arkadia	90	–		
cArkadia	83	83	–	
XArkadia	73	73	77	–

The percentage amino-acid sequence identity between mouse *Arkadia* (GenBank accession no. AF330197) and the most closely related gene products from human (Hs *Arkadia*: GenBank accession no. NT_024655), chick (c*Arkadia*: GenBank accession no. AF329846) and *Xenopus* (X*Arkadia*: GenBank accession no. AF329847) as determined by the BESTFIT algorithm of GCG programs. The RING-H2 domain (amino acids 937–977) is 100% identical in all four species.

anterior primitive streak (APS) by examining the expression of *Hnf3β* (ref. 18). As *HNF3β* is required cell-autonomously for node and mesendoderm differentiation, the severe deficit of cells expressing *Hnf3β* in the APS of *arkadia* embryos (Fig. 4a) indicates a failure to establish node precursors. In addition, we investigated the formation of endoderm in *arkadia* mutants, as this also derives from the APS. At stages before the node is morphologically visible, *Hex* and *Cer1* are normally expressed in cells that emerge from the APS to form the anterior definitive endoderm (ADE), which differentiates mainly into foregut^{13,19}. In *arkadia* mutants, expression of *Hex* (Fig. 4e) and *Cer1* (Fig. 4c, d) is markedly reduced in the ADE. As the expression of the transcription factor gene *Hex* is shown to be required cell-autonomously for the formation of ADE¹⁹, this tissue must be absent or severely reduced in *arkadia* embryos.

The absence of ADE and foregut (see *Shh* expression, Fig. 2c) in *arkadia* mutants is consistent with studies on *Hnf3β*-null embryos, which also lack both tissues. However, *Hnf3β* mutants also lack the midgut⁴, which is present in *arkadia* embryos (Fig. 2c). It is therefore likely that the few remaining cells expressing *Hnf3β* in the *arkadia* embryos contribute to midgut. From the above data we conclude that *Arkadia* acts upstream of *HNF3β* to specify node

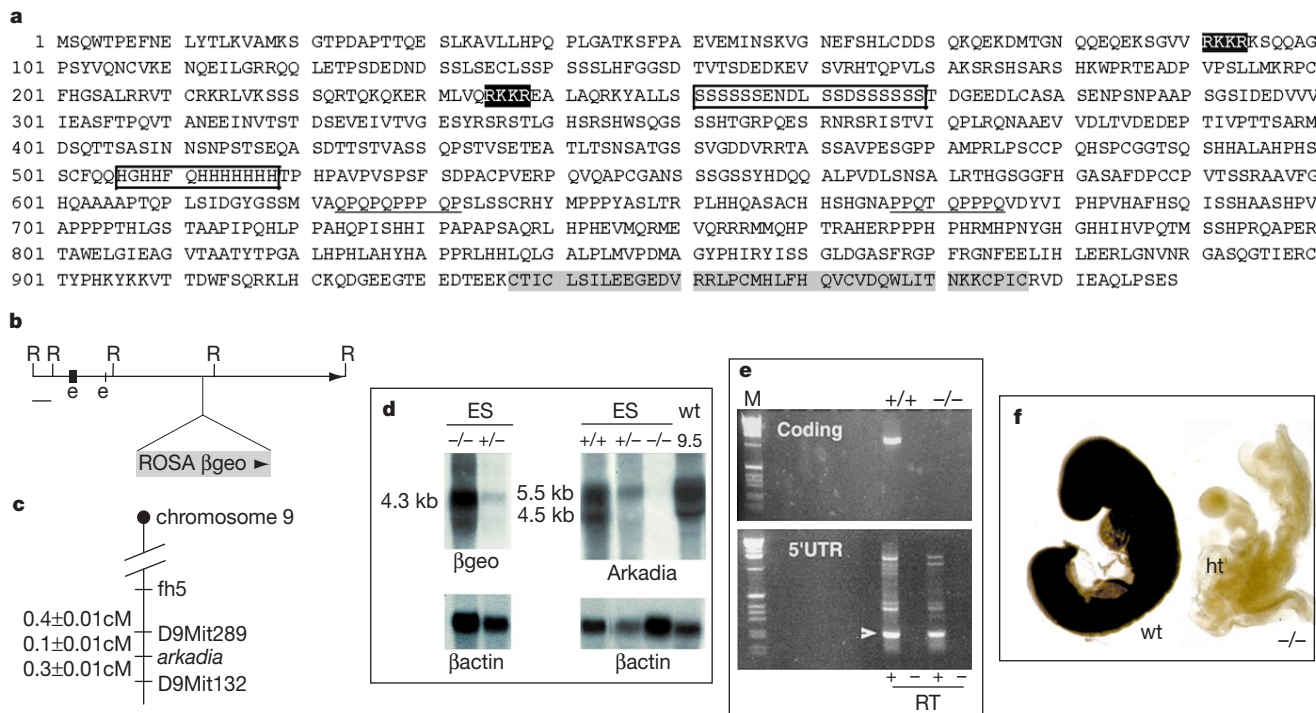


Figure 3 Molecular analysis of *Arkadia*. **a**, Predicted protein sequence of *Arkadia*. Putative nuclear localization signals highlighted black, serine-rich and histidine-rich sequences boxed, proline-glutamine repeats underlined and RING-H2 Zn finger shaded grey. **b**, *Arkadia* GT insertion. R, *EcoRI* site; e, exon; arrowheads, transcriptional orientation; bar = 1 kb. **c**, Genetic map of *arkadia* locus. cM, centimorgans. **d**, Northern blots of ES cell and mouse embryo RNA, probed as indicated below each panel. **e**, RT–PCR with *Arkadia* coding region nested primers amplifies a 1.2-kb product from wild-type

(+/-), but not -/- 8.5-d.p.c. embryo RNA. Primers from *Arkadia* 5' UTR exons (e in **b**) amplify a specific product (arrowheads) in +/+ and +/- samples. Larger products may correspond to alternatively spliced 5' UTR exons. Reverse transcriptase (RT) was omitted in control samples (labelled -). M, DNA size marker. **f**, WISH with *Arkadia* coding region probe on 9.5-d.p.c. embryos. ht, heart; -/-, *arkadia* homozygous mutant; wt, wild type (*Arkadia* +/+ or +/-).

precursors including the ADE.

Cer1 and transcription factor genes including *Hex*^{13,19}, *Otx2* and *Lim1* (refs 20, 21) are expressed not only in the ADE and node derivatives, but also in the extraembryonic signalling centre, the AVE. There is evidence that both domains act synergistically in anterior development^{22,23}. The expression of both *Hex* (Fig. 4f) and *Cer1* (Fig. 4d) in early streak *arkadia* embryos indicates that the AVE is present and correctly localized. Therefore, the AVE is probably responsible for the initiation of anterior development.

In wild-type embryos, the visceral endoderm is displaced by definitive endoderm and the AVE is replaced by the ADE¹⁹. In *arkadia* embryos, during late streak stages, expression of *Hex* and *Cer1* in the AVE disappears (Fig. 4e, c), indicating that the AVE may be displaced. This was confirmed by examining the expression of the general visceral endoderm markers *Gata4* (ref. 24) (Fig. 4b) and α -foetoprotein²⁵ (data not shown). The visceral endoderm in *arkadia* embryos must be displaced by definitive endoderm other than ADE. We conclude that in *arkadia* embryos, the combination of visceral endoderm and AVE displacement, in the absence of ADE and node, results in the degeneration of anterior structures. It would be interesting to see whether *Hnf3 β* mutant embryos, which make less definitive endoderm than *arkadia* embryos and fail to displace the AVE⁴, maintain anterior development better than *arkadia* mutants.

The *Arkadia* gene is expressed in both embryonic and extraembryonic lineages. To identify tissues in which *Arkadia* function is required to establish node precursor cells we generated mouse chimaeras. The extraembryonic lineages of the mammalian embryo do not contribute to the embryo proper (epiblast). When ES cells are injected into blastocysts they can contribute with high efficiency (up to 95% of cells) to embryonic but rarely to extraembryonic lineages²⁶.

Expression of the β geo marker gene of the GT vector in the *arkadia* locus mimics that of *Arkadia* itself. Therefore, X-gal staining was used to assess the contribution of *arkadia* mutant cells in the chimaeras. In chimaeric embryos generated with wild-type host blastocysts and *arkadia* mutant ($-/-$) ES cells all extraembryonic

lineages are exclusively wild type. However, the embryonic lineages consisted of a mixture of wild-type and ES-derived mutant cells. At 9.5 d.p.c., chimaeras with embryonic tissues comprising more than 90% *Arkadia*^{-/-} cells ($n = 10$) were phenotypically normal (Fig. 5a). Although we cannot exclude the possibility that the presence of a small number of wild-type cells (< 10%) in the embryonic lineages can rescue the *arkadia* mutant phenotype, it is likely that wild-type extraembryonic lineages are responsible for rescuing the defect.

To exclude a role for wild-type cells, we generated chimaeras in which all embryonic lineages were derived exclusively from the *arkadia* mutant ES cells and all extraembryonic lineages were wild-type. Such chimaeras were generated using tetraploid wild-type embryos aggregated with mutant ES cells. The tetraploid cells readily contribute to extraembryonic lineages but are rapidly competed out of the embryonic lineages by the diploid cells⁴. All chimaeric embryos ($n = 6$) developed a normal head and mesoderm (Fig. 5b, c) indicating that *arkadia* mutant cells can differentiate into mesoderm and anterior neuroectoderm. Therefore, *Arkadia* function is required in the extraembryonic lineages, and its action is extrinsic to the lineages it specifies.

The time at which the wild-type tetraploid cells are competed out from the early embryonic lineage (epiblast) is unknown. Consequently the transient presence of wild-type tetraploid cells in the epiblast rather than the wild-type extraembryonic lineages may have been responsible for rescuing the anterior defect. To exclude this possibility we generated reciprocal chimaeras where all extra-

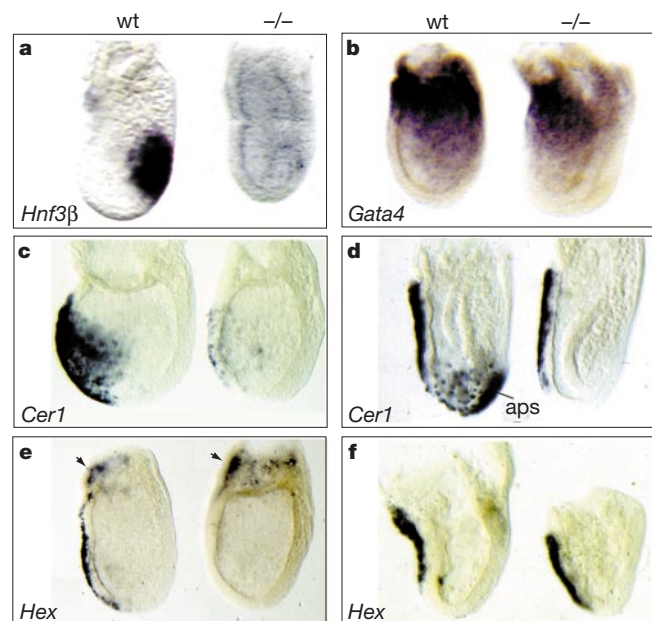


Figure 4 Anterior primitive streak (APS) defects in *arkadia* embryos. WISH on early streak (a, f), mid-streak (b, d) and NP (c, e) stage embryos probed as indicated, anterior to the left. a, HNF3 β -expressing node precursor cells are severely depleted in *arkadia* APS. b, f, Mutant embryos show normal expression of VE and AVE markers *Gata4* and *Hex*. In the mutant, *Cer1* expression is absent from distal ADE (c) and APS (d) but present in the AVE (e); similarly, *Hex* is absent from ADE (e) but present in the extraembryonic region (arrowheads) and the AVE (f). $-/-$, *arkadia* homozygous mutant; wt, wild type (*Arkadia* $+/+$ or $+/-$)

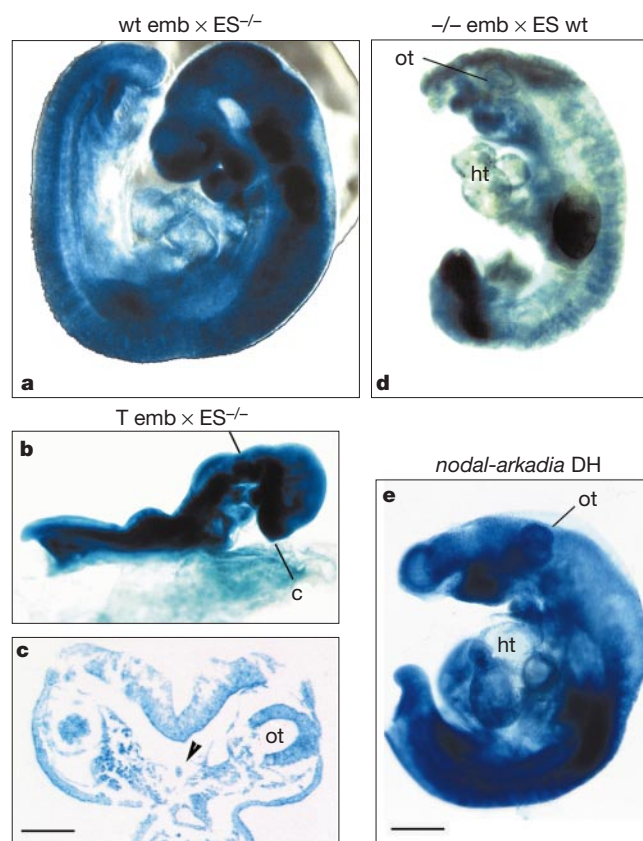


Figure 5 *Arkadia* function in the extraembryonic lineages. a, X-gal stained chimaeric embryo ($-/-$, D12/N ES cells in wild-type blastocyst) exhibiting normal appearance, despite > 90% mutant (blue) cells. b, Embryo from aggregation of $-/-$ D19 ES cells with wild-type tetraploid morulae (T emb), therefore entirely mutant, showing rescued head phenotype (anterior to the right). c, Section through the embryo in b as indicated, showing blue staining of all tissues including the notochord (arrowhead). d, Chimaeric embryo (wild-type, CCE ES cells in $-/-$ blastocyst) with anterior truncation. e, *Arkadia*^{+/+} *Nodal*^{+/+} double heterozygous embryo, exhibiting anterior truncation. DH, double heterozygous; ht, heart; ot, otic vesicle; $-/-$, *arkadia* homozygous mutant; wt, wild type (*Arkadia*^{+/+}); bars, 0.2 mm.

embryonic lineages are mutant for Arkadia but embryonic lineages derive from a mixture of mutant and wild-type cells. Such chimaeras were generated by injection of wild-type ES cells into blastocysts obtained from mating *Arkadia* heterozygotes. Only a quarter of these blastocysts were homozygous mutant, and only such chimaeric embryos ($n = 20$) showed anterior truncation at 9.5 d.p.c. (Fig. 5d). These truncations were similar to or less severe than those seen in *arkadia* mutant embryos.

The studies of chimaeras indicate that the *arkadia* phenotype depends on the genotype of the extraembryonic lineages, and that Arkadia function outside the epiblast is critical for mesendoderm and head development. We therefore conclude that the function of Arkadia to establish the progenitors of the ADE and the node is dependent on its expression within an extraembryonic cell type.

As Arkadia is unlikely to be a signal, another molecule produced by the extraembryonic lineages is likely to act as a downstream mediator of Arkadia function. Studies in other vertebrates show that the signalling factor Nodal is essential for mesendoderm induction²⁷. In mice, analysis with chimaeras shows that *Nodal* expression in the visceral endoderm is required not only for anterior development⁷ but also for mesendoderm formation (ref. 7 and E. J. Robertson, personal communication). To test whether Nodal is a secreted factor acting downstream of Arkadia in node induction, we generated double heterozygotes by mating heterozygous *Nodal*²⁸ and *Arkadia* mice. Mice heterozygous for either *nodal* or *arkadia* are always normal, irrespective of genetic background (either outbred or 129Sv/Ev). However, morphological examination (Fig. 5e) and serial sectioning (data not shown) showed that some double-heterozygote embryos exhibit variable head truncations extending to the hindbrain, loss of the midline in hindbrain tissue and occasional fusion of anterior somites across the midline. Genotyping of affected embryos confirmed, without exception, that they were double heterozygotes. However, only 10% of the double heterozygotes (2 out of 20) displayed severe head truncations. Consistent with this observation, genotyping of live-born offspring from these crosses showed that some double heterozygotes were viable and fertile. The variability in the double-heterozygote phenotype probably reflects the outbred genetic background. These results indicate that Arkadia and Nodal are probably involved in the same signalling process and support the hypothesis that Arkadia acts through Nodal to induce node-derived tissues. Parallel experiments conducted in the frog provide clear evidence that Arkadia modulates Nodal signalling¹⁷.

Emerging evidence has shown that the extraembryonic lineages, and in particular the visceral endoderm, are important in early embryonic patterning. This has been attributed mainly to the AVE and its function in anterior–posterior axis specification. Here we show that the extraembryonic lineages have an additional role: the establishment of the node, for which the Arkadia is essential. We show that this node-inducing activity is independent of the establishment and function of the AVE. However, through the requirement of the node for maintenance and further patterning of anterior embryonic structures, the node-inducing activity, like the AVE, is essential for specification of the anterior–posterior axis. An early signalling domain involved in organizer induction also exists in amphibian embryos. This is referred to as the Nieuwkoop centre and is located at the dorsal side of the blastula stage embryo adjacent to the region where Spemann's organizer forms²⁹. Convergence of two signalling pathways consisting of the dorsal modifier and the mesendoderm-inducing pathway, activated by Wnt and TGF β signals respectively, is believed to be responsible for organizer induction. The high degree of homology of the *Arkadia* gene between amniotes and anamniotes, along with experiments in frog embryos¹⁷, suggest that Arkadia function is conserved. As Arkadia influences Nodal signalling, we propose that it is a component of the mesoendodermal-inducing pathway of the vertebrate organizer-inducing centre. □

Methods

Generation of mouse strains and chimaeras

ES cell derivation, culture, and retroviral transduction were as described^{27,28}. The *arkadia* mutation was repeatedly backcrossed (> 12 generations) onto the 129Sv/Ev inbred genetic background to improve phenotypic consistency. Chimaeric embryos were generated by blastocyst injection or tetraploid morula aggregation as described^{4,7} with CCE or *arkadia*^{-/-} ES cell lines D12/N or D19. *Arkadia*^{+/-}*Nodal*^{+/-} double heterozygous embryos were generated by mating 129Sv *Nodal*^{+/-} males²⁸ and F₁(129Sv/MF1) *Arkadia*^{+/-} females.

Arkadia gene cloning

A genomic library was constructed from partially *Sau*3AI-digested *arkadia* heterozygous DNA ligated into *Xho*I-digested lambda FIXII, and packaged with Gigapack II (Stratagene). Clones spanning a region from 15 kb upstream to 6.5 kb downstream of the GT insertion were recovered by screening with a 2-kb *Eco*RI–*Eco*RV probe from the β geo reporter. A polymorphic (CA)_n repeat located 5' of the GT insertion was isolated and mapped in the European Collaborative Interspecific Backcross (EUCIB); see <http://www.hgmp.mrc.ac.uk/MBx/MBxChromosomes.html>. 5' RACE PCR was performed on cDNA from heterozygous *Arkadia* adult liver using the Marathon cDNA Amplification Kit (Clontech), with the supplied adapter primer 1 (AP1) and a *lacZ* specific primer (CTT CTG GTG CCG GAA ACC AGG C). The product was used to isolate downstream cDNA fragments from an 8.5 d.p.c. lambda gt10 library (a gift from R. Lovell-Badge). These were used to isolate further fragments from a 10.5 d.p.c. lambda Zap cDNA library (a gift from E. Robertson). The mouse cDNAs were used to isolate chick and *Xenopus* clones from embryonic chick and *Xenopus* stage 35 lambda Zap cDNA libraries (gifts from M. Jones and D. Wilkinson, respectively). The human homologue of *Arkadia* was identified by mining the human subset of dbEST at the NCBI with the mouse *Arkadia* cDNA sequence.

Arkadia gene transcription

Northern blots of total RNA, extracted with RNAzol B (Biogenesis), were hybridized in ExpressHyb hybridization solution (Clontech), washed with 0.5 $\times$ SSC, 0.1% SDS at 55 °C and exposed at -70 °C. The *Arkadia* coding region probe used for northern and WISH was RT-PCR was synthesized from total RNA of 8.5 d.p.c. *arkadia* (-/-) and wild-type (+/+) embryos. Coding region specific primer pairs for nested RT-PCR were AkdF (5' CTCACAAGTGGCTCGAACT 3' nucleotides 884–903) and AkdR (5' AGATGTGA-CAGGGCAGCAG 3' 2,124–2,106), followed by AkdNesF (5' GCAGATCCTGTGC-CAAGTTT 3' 907–926) and AkdNesR (5' AAGCAGATGCACCATGGAA 3' 2,098–2,080). Primers to detect transcription upstream of the GT insertion were 5' UTRxon1F (5' GGCTCAGCCTCAGCTCC 3' 116–134) and 5' UTRxon2R (5' TTCTCCGGAA-CAAAAACCGA 3' 320–301).

Embryo genotyping, staging and WISH

Adult *Nodal* and *Arkadia* heterozygotes were identified by PCR amplification of genomic DNA or X-gal staining of tail or pinna biopsies (see below). Embryos were genotyped by PCR amplification of DNA from extraembryonic tissue (pre-somite and early somite stages) or yolk sac endoderm (late somite stage chimaeras). Oligonucleotide primer pairs TGAGGTAGGCATACCTAGAG and TGACTTAAGCCCTGCAATCC; TGAGGTAGGCA-TACCTAGAG and CTGAGTGATTGACTACCCGT; and TCTGGATTCATCGACTGTGG and CTGGATGTAGGCATGGTGGTAG were used to give diagnostic amplification products of 313 bp for the wild-type *Arkadia* allele, 293 bp for disrupted *Arkadia* and 925 bp for disrupted *Nodal*, respectively. Pre-somite embryos were staged as described³⁰. Embryos shown in Figs 1c, f–i, 2c–e, 3g and 4c–f were from early (<7) backcross generations. WISH of embryos with digoxigenin-labelled cRNA probes was performed as described⁷. Embryos were photographed in pairs to allow size comparisons, on Kodak 64T or 160T film. Some images were inverted during compilation of the Figures to simplify annotation.

Received 1 November 2000; accepted 16 January 2001.

- Beddington, R. S. & Robertson, E. J. Anterior patterning in mouse. *Trends Genet.* **14**, 277–284 (1998).
- Beddington, R. S. & Robertson, E. J. Axis development and early asymmetry in mammals. *Cell* **96**, 195–209 (1999).
- Tam, P. P. & Behringer, R. R. Mouse gastrulation: the formation of a mammalian body plan. *Mech. Dev.* **68**, 3–25 (1997).
- Dufort, D., Schwartz, L., Harpal, K. & Rossant, J. The transcription factor HNF3beta is required in visceral endoderm for normal primitive streak morphogenesis. *Development* **125**, 3015–3025 (1998).
- Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L. & Kuehn, M. R. Nodal is a novel TGF-beta-like gene expressed in the mouse node during gastrulation. *Nature* **361**, 543–547 (1993).
- Conlon, F. L., Barth, K. S. & Robertson, E. J. A novel retrovirally induced embryonic lethal mutation in the mouse: assessment of the developmental fate of embryonic stem cells homozygous for the 413.d proviral integration. *Development* **111**, 969–981 (1991).
- Varlet, L., Collignon, J. & Robertson, E. J. nodal expression in the primitive endoderm is required for specification of the anterior axis during mouse gastrulation. *Development* **124**, 1033–1044 (1997).
- Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5**, 1513–1523 (1991).
- Wood, H. B. & Episkopou, V. Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mech. Dev.* **86**, 197–201 (1999).
- Rowitch, D. H. & McMahon, A. P. Pax-2 expression in the murine neural plate precedes and encompasses the expression domains of Wnt-1 and En-1. *Mech. Dev.* **52**, 3–8 (1995).
- Dressler, G. R., Deutsch, U., Chowdhury, K., Nornes, H. O. & Gruss, P. Pax2, a new murine paired-

- box-containing gene and its expression in the developing excretory system. *Development* **109**, 787–795 (1990).
12. Albano, R. M. & Smith, J. C. Follistatin expression in ES and F9 cells and in preimplantation mouse embryos. *Int. J. Dev. Biol.* **38**, 543–547 (1994).
 13. Belo, J. A. *et al.* Cerberus-like is a secreted factor with neutralizing activity expressed in the anterior primitive endoderm of the mouse gastrula. *Mech. Dev.* **68**, 45–57 (1997).
 14. Echelard, Y. *et al.* Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. *Cell* **75**, 1417–1430 (1993).
 15. Wilkinson, D. G., Bhatt, S. & Herrmann, B. G. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343**, 657–659 (1990).
 16. Loric, K. L. *et al.* RING fingers mediate ubiquitin-conjugating enzyme (E2)-dependent ubiquitination. *Proc. Natl Acad. Sci. USA* **96**, 11364–11369 (1999).
 17. Neiderländer C., Walsh, J. L., Episkopou, V. & Jones, C. M. Arkadia enhances nodal-related signalling to induce mesendoderm. *Nature* **410**, 830–834 (2001).
 18. Filosa, S. *et al.* Goosecoid and HNF-3beta genetically interact to regulate neural tube patterning during mouse embryogenesis. *Development* **124**, 2843–2854 (1997).
 19. Martínez Barbera, J. P. *et al.* The homeobox gene Hex is required in definitive endodermal tissues for normal forebrain, liver and thyroid formation. *Development* **127**, 2433–2445 (2000).
 20. Rhinn, M. *et al.* Sequential roles for Otx2 in visceral endoderm and neuroectoderm for forebrain and midbrain induction and specification. *Development* **125**, 845–856 (1998).
 21. Shawlot, W. *et al.* Lim1 is required in both primitive streak-derived tissues and visceral endoderm for head formation in the mouse. *Development* **126**, 4925–4932 (1999).
 22. Tam, P. P. & Steiner, K. A. Anterior patterning by synergistic activity of the early gastrula organizer and the anterior germ layer tissues of the mouse embryo. *Development* **126**, 5171–5179 (1999).
 23. Bachiller, D. *et al.* The organizer factors Chordin and Noggin are required for mouse forebrain development. *Nature* **403**, 658–661 (2000).
 24. Arceci, R. J., King, A. A., Simon, M. C., Orkin, S. H. & Wilson, D. B. Mouse GATA-4: a retinoic acid-inducible GATA-binding transcription factor expressed in endodermally derived tissues and heart. *Mol. Cell. Biol.* **13**, 2235–2246 (1993).
 25. Dziadek, M. & Adamson, E. Localization and synthesis of alphafoetoprotein in post-implantation mouse embryos. *J. Embryol. Exp. Morphol.* **43**, 289–313 (1978).
 26. Beddington, R. S. & Robertson, E. J. An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo. *Development* **105**, 733–737 (1989).
 27. Schier, A. F. & Shen, M. M. Nodal signalling in vertebrate development. *Nature* **403**, 385–389 (2000).
 28. Collignon, J., Varlet, I. & Robertson, E. J. Relationship between asymmetric nodal expression and the direction of embryonic turning. *Nature* **381**, 155–158 (1996).
 29. Harland, R. & Gerhart, J. Formation and function of Spemann's organizer. *Annu. Rev. Cell Dev. Biol.* **13**, 611–667 (1997).
 30. Downs, K. M. & Davies, T. Staging of gastrulating mouse embryos by morphological landmarks in the dissecting microscope. *Development* **118**, 1255–1266 (1993).

Acknowledgements

Gene trap mutagenesis was done in the laboratory of E. J. Robertson at Columbia University, New York. We thank R. Lovell-Badge, T. Jessell, E. Robertson and A. Streit for reading the manuscript; E. L. Ferguson for discussions; E. Robertson for the gift of the GT-carrier and *nodal* (+/–) mice and T. Rodriguez for rederivation of the latter; P. Soriano for the *Rosa* β geo vector and virus-producing cell lines; S.-L. Ang, R. Beddington, E. De Robertis, P. Gruss, T. Jessell and J. Rossant for WISH probes; M. Alexiou, H. Wood and S. Malas for advice and support; and A. Simpkins, R. Grahame, G. Baker and E. Damien for technical support.

Correspondence and requests for materials should be addressed to V.E.
(e-mail: vasso.episkopou@csc.mrc.ac.uk).

Arkadia enhances nodal-related signalling to induce mesendoderm

Christiane Niederländer^{*}, James J. Walsh[†], Vasso Episkopou[†] & C. Michael Jones^{*}

^{*} Section of Gene Function and Regulation, Chester Beatty Laboratories, Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK

[†] Mammalian Neurogenesis Group, MRC Clinical Sciences Centre, Imperial College School of Medicine, Hammersmith Hospital, DuCane Road, London W12 0NN, UK

Nodal-related members of the transforming growth factor (TGF)- β family regulate the induction of mesoderm, endoderm, and mesendoderm, a tissue specific to the Spemann organizer^{1–7}. How these different tissues form in response to the same signalling molecules is not completely understood. It has been suggested that concentration-dependent effects, mediated by extracellular

cofactors and antagonists, are responsible for the differences^{1,8–10}. Here we show that the nuclear protein Arkadia specifically potentiates the mesendoderm-inducing activity of a subset of TGF- β family members. The combined activities of Arkadia and *Xenopus* nodal-related-1 are sufficient to induce mesendoderm and suppress mesoderm. Arkadia dorsalizes ventral tissues, resulting in the induction of organizer-specific gene expression. Blocking nodal signalling extracellularly inhibits these effects. Arkadia influences nodal activity when co-expressed and can function in cells adjacent to those producing the nodal signal. Our findings, together with the observation that Arkadia mutant mice lack a node and node-derived mesendoderm, identify Arkadia as an essential modulator of the nodal signalling cascade that leads to induction of Spemann's organizer.

Arkadia is essential for establishing the mammalian organizer. Mice null for *Arkadia* fail to develop a prechordal plate and notochord, and subsequently exhibit anterior truncations¹¹. To investigate the function of Arkadia and its role in establishing Spemann's organizer, we have isolated the *Xenopus* cognate and studied its activities in *Xenopus* embryos. Both the mouse and *Xenopus* complementary DNAs exhibit identical activities. *Arkadia* transcripts are encoded maternally and persist into late tadpole stages of development (Fig. 1a). Whole-mount *in situ* hybridization detected ubiquitous maternal expression in fertilized eggs (data not shown), with the highest transcript levels on the presumptive dorsal side of blastulae, where Spemann's organizer forms (Fig. 1b). At gastrula stages, expression was detected throughout the marginal zone and endoderm (Fig. 1c–e) but was not detected in the ectoderm (Fig. 1e).

Misexpression of Arkadia RNA on the ventral side of the embryo led to the dorsalization of ventral marginal zone tissues (VMZs) (Fig. 1f–k); explants underwent convergent extension movements (Fig. 1g–k) and developed muscle (Fig. 1, insets). Arkadia-injected VMZs expressed the Spemann-organizer-specific markers *Otx2* and *goosecoid*; the highest levels of Arkadia induced the anterior endodermal markers *Xhex* and *cerberus* (Fig. 1l, lanes 3–5). Thus, increasing the levels of Arkadia dorso-anteriorized both mesodermal and endodermal tissues.

To test if this 'dorsalization' of VMZs resulted from the inhibition of endogenous ventralizing signals mediated by bone morphogenetic proteins (BMPs)^{12–16}, we co-expressed BMP4 and Arkadia in animal caps. The expression of *Xbra*, a mesodermal marker¹⁷, and *Xhox3*, a marker of BMP4 signalling¹⁸ was identical in animal caps injected with BMP4/Arkadia or BMP4 alone (Fig. 1m, compare lane 10 and 11). Furthermore, expression of Arkadia in animal caps did not induce mesoderm (Fig. 1m, lane 9), endoderm (see Fig. 2n, lane 2) or neural tissue (data not shown). Hence, Arkadia does not inhibit signalling by BMPs, or exhibit any inducing activity on its own.

These results suggested that Arkadia activity might depend on other factors, and if so, these should be present in vegetal and marginal regions of the embryo, where the Arkadia-dorsalizing effects were observed. To test this idea, we used an animal cap assay in which TGF- β factors that are expressed in the vegetal hemisphere induce mesoderm and/or endoderm/mesendoderm at low and high concentrations, respectively^{19,20}. When co-injected as RNAs, Arkadia enhanced the signalling properties of Activin, BVg (an active Vg1), *Xnr1* and *Xnr2* (Fig. 2). When low levels of the inducing factors were used in combination with Arkadia, the tissues and molecular markers induced were indicative of a higher concentration of inducing factor (Fig. 2m). When compared with those injected with inducer alone, co-injected explants either underwent more marked convergent extension movements (Fig. 2e, f) or exhibited a rounded 'inside-out' appearance indicative of endoderm formation²¹ (Fig. 2g, h). In combination with *Xnr1*, mesoderm induction (*Xbra* and actin expression) was completely suppressed in favour of endoderm formation (*Sox17* β expression;

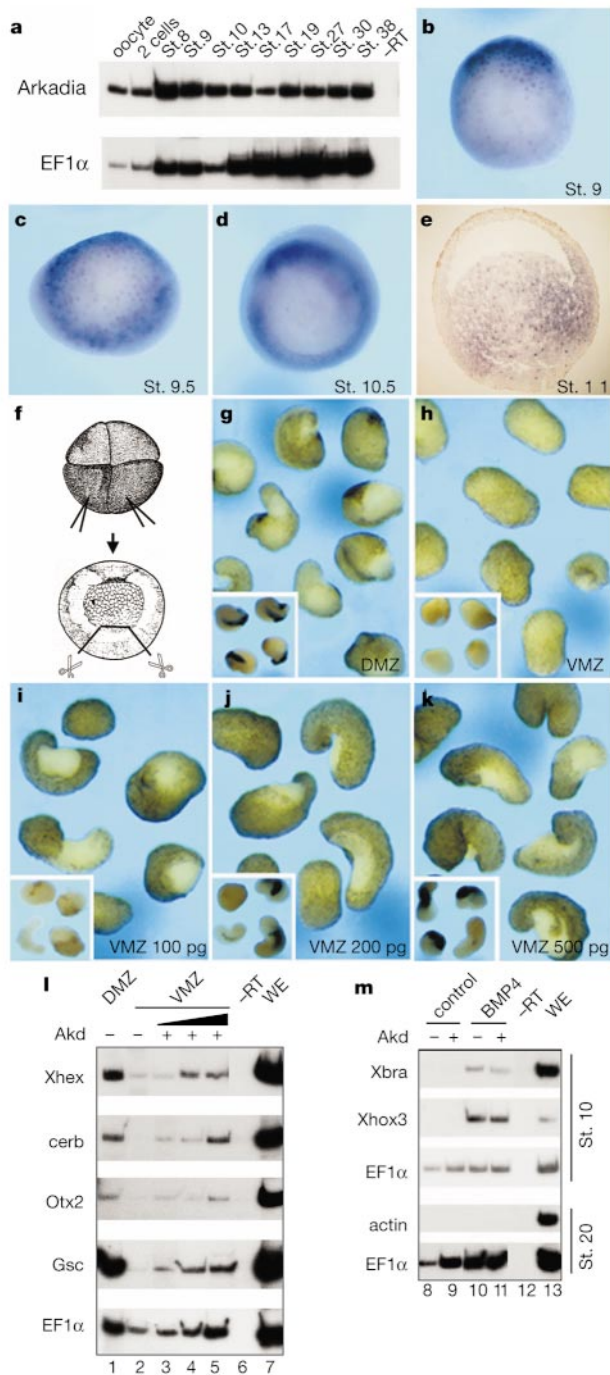


Figure 1 Arkadia expression and dorsalizing activity in ventral marginal zone explants (VMZs). **a**, RT-PCR detected Arkadia transcripts maternally and throughout development. **b–e**, Highest expression is detected by *in situ* hybridization on the presumptive dorsal side of mid blastulae (**b**) and spreads throughout the endoderm and marginal zone by mid gastrula stages (**c–e**). Expression is most evident in a tangential section at stage 11 (**e**). Whole-mount pictures are vegetal views, dorsal to the top. Transcripts were detected in the ectoderm by RT-PCR (data not shown). **f–i**, The effect of Arkadia was tested by injecting RNA into ventral blastomeres and subsequently isolating VMZ explants at early gastrula stages (**f**, **g**, **h**). In contrast to VMZ controls (**h**), dorsal marginal zones (DMZs) elongate (**g**) and stain positively for muscle with antibody 12/101 (inset). **i–k**, VMZs injected with increasing concentrations of Arkadia RNA elongate progressively more, and stain positively for muscle (insets, $n \geq 10$). **l**, Arkadia-induced expression of dorsal markers Otx2 and Gsc and the anterior endodermal markers Xhex and cerberus (lanes 3–5). **m**, Arkadia neither inhibited BMP signalling nor exhibited any inducing activities in animal cap assays (lane 9; see also Fig. 2). Whole embryo RNA (WE) served as a positive control for RT-PCR; a similar sample minus reverse transcriptase (–RT) in the cDNA reaction served as a negative control (**l**, **m**).

Fig. 2m, compare lanes 7 and 8). Arkadia interacted differently with Xnr1 and Xnr2, the most closely related family members, consistent with previous reports of differences between them^{22,23}.

In vertebrates, nodal signalling is essential for the formation of endoderm and mesoderm^{3,4,20,24}—the two germ layers in which Arkadia activity was observed. Injection of Arkadia RNA into one ventro-vegetal blastomere of 8–16-cell embryos induced partial secondary axes, similar to those induced by Xnr1 (Fig. 3a, b). Arkadia-induced secondary axes contained segmented somites but no notochord, and expressed the spinal cord marker HoxB9 in dorsal posterior regions (Fig. 3c, d, see also Supplementary Information). Tissue was present anterior to HoxB9 expression, but more anterior markers (Krox20, Otx2) were not detected (data not shown). Overexpression of Arkadia throughout the embryo resulted in a hyperdorsalized phenotype and a severe disruption of gastrulation, again similar to results obtained by ubiquitous overexpression of nodal-related proteins (C.M.J., unpublished data). The specificity of Arkadia's activities was underlined by the inability of Arkadia to alter inductions mediated by *derrière*, a TGF- β

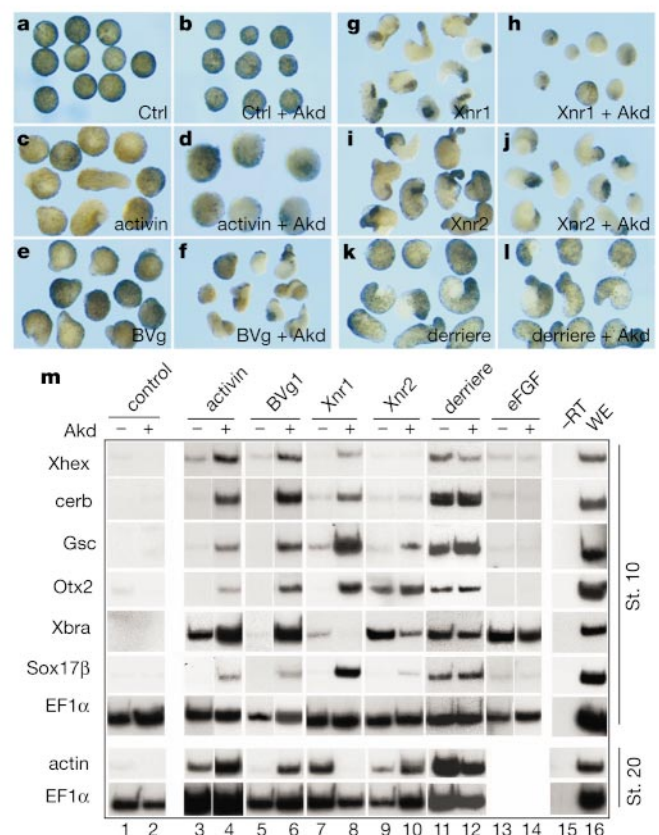


Figure 2 Arkadia enhanced the activities of dorsalizing TGF- β factors **a–l**. Morphology of explants after injection of inducer $\pm$ Arkadia RNA (250 pg). Inducing factors were used at limiting concentrations to highlight the effects of Arkadia. **a**, **b**, Arkadia-injected explants (250 pg) (**b**) are indistinguishable from controls (**a**). **c**, **d**, Weak activin (5 pg) inductions (**c**) were accentuated by Arkadia (**d**). Explants acquired an 'inside-out' endodermal appearance²¹. **e**, **f**, BVg-induced (25 pg) explants (**e**) elongated more markedly and acquired a more dorsal character when co-expressing Arkadia (**f**). **g**, **h**, Xnr1-induced (100 pg) effects (**g**) were enhanced such that explants acquired an endodermal morphology (**h**). **i**, **j**, Arkadia augmented Xnr2 (100 pg) effects only slightly. **k**, **l**, Inductions by *derrière* were unaffected by Arkadia. **m**, Arkadia enhanced the expression of mesodermal markers in the presence of dorsalizing TGF- β factors. Arkadia alone induced none of the molecular markers (lane 2). Strongest effects are observed in combination with Xnr1. Mesoderm induction (Xbra and actin expression) was suppressed in lieu of anterior mesoderm (compare lanes 7,8). Arkadia had no effect on inductions by *derrière* (lanes 11,12) or eGFG (lanes 13,14).

factor closely related to Vg1 but essential for the development of posterior tissues²⁵, or the different signalling molecule embryonic fibroblast growth factor (eFGF) (Fig. 2). These results suggest that Arkadia modulates signalling of the specific subset of TGF- β factors responsible for formation of dorso-anterior tissues of Spemann's organizer. As Arkadia enhanced endogenous nodal signalling, its effects on the ventral side of the embryo were similar to those of nodal-related factors.

We next determined whether Arkadia function requires nodal signalling by using a specific extracellular nodal-antagonist, Cer-S¹⁴.

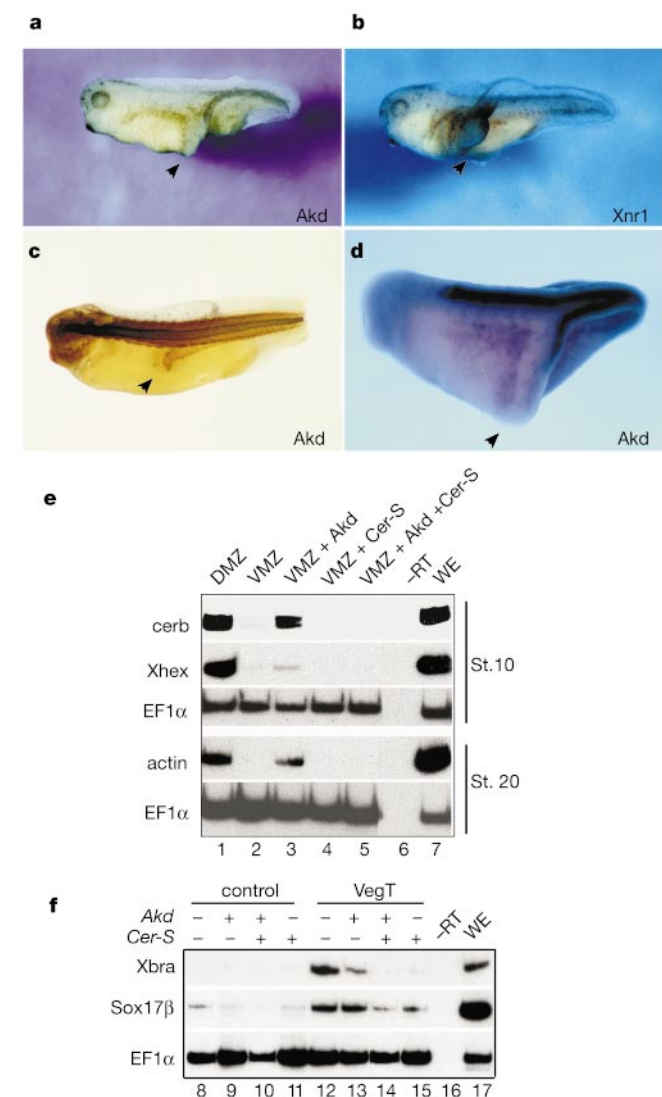


Figure 3 Arkadia mimicked Xnr1 effects and required nodal signalling to function. **a–d**, Secondary axes resulting from ventral Arkadia injections (**a**; 500 pg, 15% of injected embryos) are similar to those formed in response to Xnr1 (**b**; 50 pg, 50% of injected embryos). **c, d**, Arkadia-induced secondary axes contained segmented somites (antibody 12/101, brown), but no notochord (antibody MZ15, indigo) (**c**), and expressed HoxB9 in posterior regions (**d**). Arrowheads delineate secondary axes. **e**, Arkadia effects were counteracted by Cer-S, a specific nodal inhibitor. VMZs were injected with 500 pg Arkadia RNA (lane 3), 100 pg Cer-S RNA (lane 4) or both (lane 5). Expression of anterior endoderm markers and induction of actin were blocked by Cer-S. **f**, Arkadia promoted endoderm formation in response to VegT. VegT-injected animal caps express mesoderm and endoderm markers (lane 12). Arkadia coexpression suppressed Xbra while Sox17 β expression remained unaffected (lane 13). In the presence of Cer-S, Xbra expression was eliminated and Sox17 β transcript levels decreased (lane 14). Note Cer-S inhibited Xbra expression induced by VegT (lane 15). WE and -RT are positive and negative controls, respectively.

The dorsalizing effects of Arkadia on VMZs were completely blocked by Cer-S. Induction of the dorsal mesodermal marker actin and the anterior endodermal markers cerberus and Xhex was inhibited (Fig. 3e), as were the morphogenetic movements associated with dorsalization (data not shown).

To analyse further the interaction between Arkadia and nodal, we tested Arkadia in combination with VegT, a vegetally localized maternal transcription factor²⁶ essential for mesoderm and endoderm formation. VegT functions by inducing the zygotic expression of TGF- β family members, particularly nodal-related factors^{19,27}. Arkadia altered the ability of VegT to induce endoderm and mesoderm in animal cap assays. The presence of Arkadia led to the suppression of mesoderm, whereas endoderm (Sox 17B expression) was sustained (Fig. 3f); results were comparable to the effects elicited by high concentrations of Xnr1. Arkadia was unable to override the general reduction of mesoderm when VegT and Cer-S were co-expressed. We conclude that Arkadia activity is critically dependent on nodal signalling, and that Arkadia functions in the cascade of inductions regulated by VegT.

Arkadia encodes a protein of 989 amino acids. The only recognizable motifs in the protein are a RING H2 domain at the carboxy-terminal end and two putative nuclear localization signals nearer the amino terminus (see ref. 11). We investigated the cellular

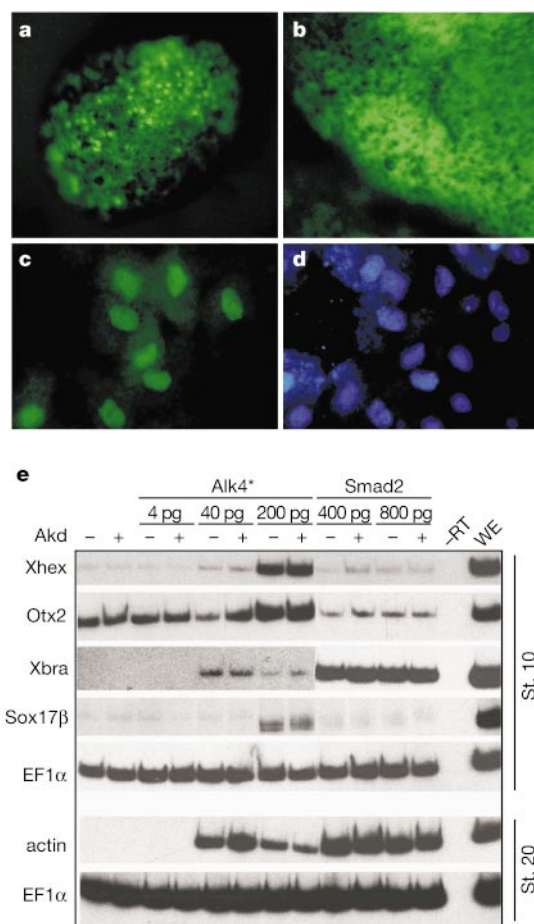


Figure 4 Arkadia does not alter signalling of TGF- β downstream effectors. **a–d**, GFP-tagged Arkadia localized to the nucleus (**a, c**), which was confirmed by co-localization with DAPI (**d**). GFP alone distributed throughout the cell (**b**). **e**, Arkadia did not potentiate mesoderm induction mediated by the TGF- β downstream effectors Alk4 or Smad2. Animal caps injected with progressively higher concentrations of the constitutively active receptor Alk4 (indicated by an asterisk) or Smad2 express mesoderm and endoderm markers. These inductions were unaffected by coinjection of Arkadia RNA (500 pg). RNA from whole embryos (WE) and -RT are positive and negative controls for RT-PCR.

localization of Arkadia using a green fluorescent protein (GFP)–Arkadia fusion protein, which localized to the nucleus when expressed in *Xenopus* embryos (Fig. 4a, c). GFP alone was detected throughout the cell (Fig. 4b). The fusion protein retained biological activity and was nuclear when co-expressed with nodal-related factors (data not shown). Therefore, nuclear localization of Arkadia–GFP was unaffected by the presence of factors with which it exerts its functions.

The nuclear localization suggested that Arkadia acted in concert with the downstream effectors of nodal-related signalling. Although the nodal receptor is not known, signalling is thought to be mediated by specific type I receptors that transduce signals through intracellular effectors, Smads²⁸. The mesendoderm inducers that Arkadia specifically affects signal through the receptor Alk4 and the effector Smad2 (refs 1, 28). In *Xenopus*, injection of these signal transduction molecules induces the same biological effects as the dorsalizing TGF- β ligands²¹. We determined whether Arkadia potentiated the effects of the constitutively active receptor Alk4 or the intracellular effector Smad2. Co-expression of Arkadia together with constitutively active Alk4 or with Smad2 did not alter the morphology of animal caps (data not shown), or lead to a difference in gene expression (Fig. 4e). Thus, the enhancing effects of Arkadia do not seem to be mediated through these downstream signalling components.

Despite this lack of cooperation with presumed nodal intracellular effectors, Arkadia might still function within cells receiving the nodal signal. Alternatively, Arkadia might modulate ligand levels or presentation and therefore operate in cells that produce nodal. To address these possibilities, we analysed the effects of localized expression of Xnr1 and/or Arkadia in the animal hemisphere of blastula embryos (Fig. 5a). We used Xnr1 because the most marked effects of Arkadia occurred in combination with this ligand (see Fig.

2). Under these conditions, Xnr1 effects could be detected within the clone of cells synthesizing Xnr1 as well as a few cell diameters outside. We assayed for mesoderm and mesendoderm induction indicative of low and high Xnr1 signalling using Xbra and Gsc as markers, respectively².

Xbra transcripts were detected around the periphery of an Xnr1-expressing clone, whereas Gsc transcripts were detected only in the cells producing the ligand, where the highest level of Xnr1 signalling occurred (Fig. 5b, c). In contrast, when Xnr1 and Arkadia were co-expressed, Xbra was suppressed, consistent with the results obtained when tested in animal cap explants (Fig. 2). In the clone expressing Xnr1 and Arkadia, Gsc transcripts became localized to the periphery, where injected and uninjected cells intermixed (Fig. 5e). Both Xnr1-injected and non-injected cells expressed Gsc. The centre of the clone, however, was devoid of Gsc transcripts. This result shows that Arkadia can function in cells producing the nodal signal, and that signalling is enhanced in the presence of Arkadia. The limited signalling range of Xnr1 made it impossible to determine whether the effects of Arkadia and Xnr1 co-expression resulted in a qualitative change in nodal signalling (modifying the specificity of the induction toward mesendoderm and away from mesoderm) or whether the effects were simply quantitative (that is, more nodal signal produced in the same limited signalling area).

The results obtained by co-expression were qualified by experiments in which Xnr1-expressing cells abutted a clone of cells expressing Arkadia. When Arkadia was expressed in cells adjacent to an Xnr1-producing clone, Xbra transcription was suppressed within the region of cell intermixing (Fig. 5f, g). Arkadia altered the response to nodal in cells adjacent to those producing the signal. This pointed to a further explanation of the co-expression results. As demonstrated by the expression of Gsc in the Xnr1-expressing clone (Fig. 5c), Xnr1-expressing cells also respond to the signal they

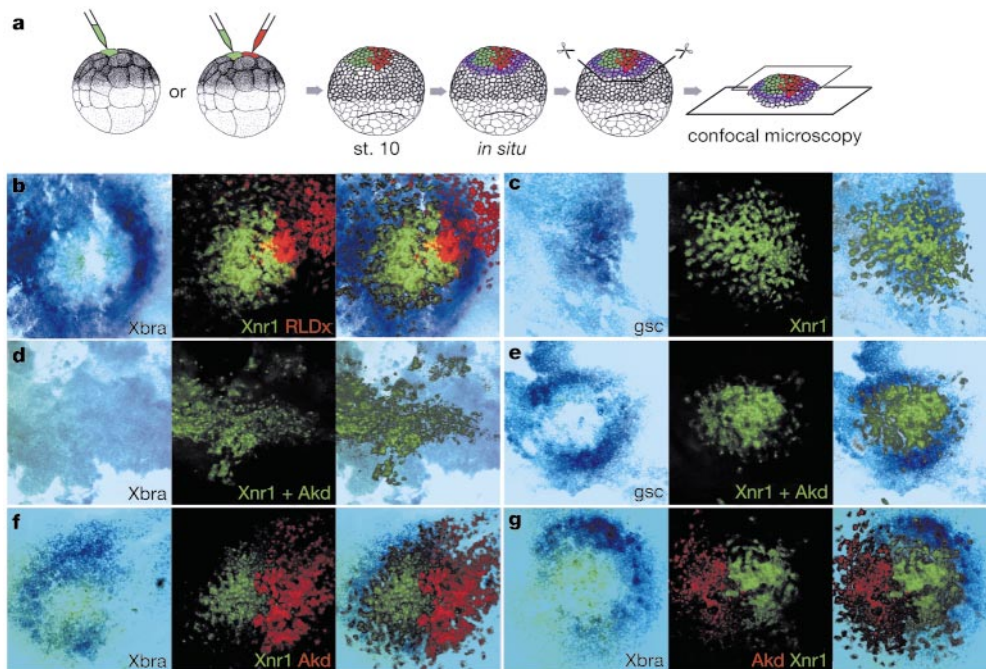


Figure 5 Arkadia functions in cells receiving the nodal signal and alters nodal signalling when co-expressed. **a**, Single animal blastomeres at the 32–64-cell stage were injected with Xnr1 RNA plus fluorescein lysinated dextran (FLDx, green), with Xnr1 and Arkadia RNA plus FLDx, or in one cell with Xnr1 plus FLDx and in an adjacent cell with rhodamine-conjugated dextran (RLDx, red) with or without Arkadia. Xbra or goosecoid expression was assayed at gastrula stages (> 20 embryos for each regimen). Assayed regions were flat-mounted in glycerol for confocal microscopy. In **b–g**, shown are a brightfield image illuminating the *in situ* hybridization signal (left), the location of the fluorescent cells

(middle), and an overlay of the two (right). **b, c**, Xbra transcripts were detected in a ring adjacent to cells expressing Xnr1 (green). RLDx was injected in an adjacent cell as a control (**b**). Gsc transcripts were detected only within the Xnr1-expressing clone (**c**). **d, e**, Co-expression of Arkadia and Xnr1 (green) led to the suppression of Xbra expression (**d**), whereas Gsc transcripts were induced at the periphery of Xnr1/Arkadia-expressing cells (**e**). **f, g**, Xbra expression was suppressed in cells expressing Arkadia (red), which were adjacent to cells expressing Xnr1 (green), indicating that Arkadia can function in cells that receive the Xnr1 signal.

produce. Hence, the effects of Arkadia and Xnr1 co-expression could be due to an increased sensitivity to the nodal signal. These effects might be mediated through the nodal positive feedback loop^{20,29}, through the upregulation of extracellular cofactors¹, or through an alternative signalling pathway.

Our results show that Arkadia modulates the activities of a subset of TGF- β factors. The Arkadia-nodal interaction leads to the induction of mesendoderm at the expense of mesoderm. This activity provides insights into how a single inductive signal can be modified to induce different tissues. Our results indicate that Arkadia can exert its effects in cells receiving the nodal signal, but it also augmented nodal signalling when co-expressed. Furthermore, Arkadia enhanced the activity of nodal-related factors induced by maternal VegT. We suggest that the initial activity of Arkadia on the dorsal side of the embryo leads to a local potentiation of nodal signalling, which results in induction of mesendoderm and Spemann's organizer. □

Methods

Cloning of *Xenopus* Arkadia

Initially, we screened a *Xenopus* stage-35 lambda Zap library at low stringency using the entire mouse cDNA. Three overlapping clones encompassing 3' sequences of *Xenopus* Arkadia were obtained and the longest was used to screen a stage-10 dorsal lip library at high stringency. The longest clone recovered ended 150 base pairs downstream from the predicted translation start.

Embryo manipulation

Embryo manipulations, *in situ* hybridization, immunostaining and β -galactose detection have been described¹⁸. The Arkadia probe consisted of 3' coding and untranslated region sequences. Injection of Arkadia RNA was performed with *in vitro* transcribed capped RNA encoding either the mouse cDNA or a hybrid cDNA encoding the first 50 amino acids of the mouse protein fused to the remainder of the *Xenopus* coding region.

RT-PCR

RNA was extracted from tissue using a guanidinium-thiocyanate buffer and RT-PCR performed as described¹⁸. The primers and cycle numbers used were according to the *Xenopus* Molecular Marker Resource (XMMR) with the exception of Arkadia, for which primers were ccacatctcacagaatgcac (forward) and ccacatctcacagaatgtgac (reverse).

Received 13 November 2000; accepted 25 January 2001.

- Schier, A. F. & Shen, M. M. Nodal signalling in vertebrate development. *Nature* **403**, 385–389 (1999).
- Jones, C. M., Kuehn, M. R., Hogan, B. L. M., Smith, J. C. & Wright, C. V. E. Nodal-related signals induce axial mesoderm and dorsalize mesoderm during gastrulation. *Development* **121**, 3651–3662 (1995).
- Conlon, F. L. *et al.* A primary requirement for nodal in the formation and maintenance of the primitive streak in the mouse. *Development* **120**, 1919–1928 (1994).
- Feldman, B. *et al.* Zebrafish organizer development and germ-layer formation require nodal-related signals. *Nature* **395**, 181–185 (1998).
- Rebagliati, M. R., Toyama, R., Fricke, C., Haffter, P. & Dawid, I. B. *cyclops* encodes a nodal-related factor involved in midline signaling. *Proc. Natl Acad. Sci. USA* **95**, 9932–9937 (1998).
- Sampath, K. *et al.* Induction of the zebrafish ventral brain and floorplate requires *cyclops/nodal* signalling. *Nature* **395**, 185–189 (1998).
- Zhou, X., Sasaki, H., Lowe, L., Hogan, B. L. M. & Kuehn, M. R. Nodal is a novel TGF- β -like gene expressed in the mouse node during gastrulation. *Nature* **361**, 543–547 (1993).
- Piccolo, S. *et al.* The head inducer Cerberus is a multifunctional antagonist of Nodal, BMP and Wnt signals. *Nature* **397**, 707–710 (1999).
- Gritsman, K. *et al.* The EGF-CFC protein one-eyed pinhead is essential for nodal signaling. *Cell* **97**, 121–132 (1999).
- Bigrove, B. W., Essner, J. J. & Yost, H. J. Regulation of midline development by antagonism of lefty and nodal signaling. *Development* **126**, 3253–3262 (1999).
- Episkopou, V. *et al.* Induction of the mammalian node requires Arkadia function in the extra-embryonic lineages. *Nature* **410**, 825–830 (2001).
- Graff, J. M., Thies, R. S., Song, J. J., Celese, A. J. & Melton, D. A. Studies with a *Xenopus* BMP receptor suggest that ventral mesoderm-inducing signals override dorsal signals *in vivo*. *Cell* **79**, 169–179 (1994).
- Suzuki, A. *et al.* A truncated bone morphogenetic protein receptor affects dorsal-ventral patterning in the early *Xenopus* embryo. *Proc. Natl Acad. Sci. USA* **91**, 10255–10259 (1994).
- Piccolo, S., Sasai, Y., Lu, B. & De Robertis, E. M. Dorsal-ventral patterning in *Xenopus*: inhibition of ventral signals by direct binding of chordin to BMP-4. *Cell* **86**, 589–598 (1996).
- Zimmerman, L. B., De Jesus-Escobar, J. M. & Harland, R. M. The Spemann organizer signal noggin binds and inactivates Bone Morphogenetic Protein 4. *Cell* **86**, 599–606 (1996).
- Dale, L. & Jones, C. M. BMP signalling in early *Xenopus* development. *BioEssays* **21**, 751–760 (1999).
- Smith, J. C., Price, B. M., Green, J. B., Weigel, D. & Herrmann, B. G. Expression of a *Xenopus* homolog of Brachyury (T) is an immediate-early response to mesoderm induction. *Cell* **67**, 79–87 (1991).
- Jones, C. M., Dale, L., Hogan, B. L., Wright, C. V. & Smith, J. C. Bone morphogenetic protein-4 (BMP-4) acts during gastrula stages to cause ventralization of *Xenopus* embryos. *Development* **122**, 1545–1554 (1996).
- Clements, D., Friday, R. V. & Woodland, H. R. Mode of action of VegT in mesoderm and endoderm formation. *Development* **126**, 4903–4911 (1999).

- Aguis, E., Oelgeschlager, M., Wessely, O., Kemp, C. & DeRobertis, E. M. Endodermal Nodal-related signals and mesoderm induction in *Xenopus*. *Development* **127**, 1173–1183 (2000).
- Armes, N. A. & Smith, J. C. The ALK-2 and ALK-4 activin receptors transduce distinct mesoderm-inducing signals during early *Xenopus* development but do not co-operate to establish thresholds. *Development* **124**, 3797–3804 (1997).
- Sampath, K., Cheng, A. M. S., Frisch, A. & Wright, C. V. E. Functional differences among *Xenopus* nodal-related genes in left-right axis determination. *Development* **124**, 3293–3302 (1997).
- Yasuo, H. & Lemaire, P. A two-step model for the fate determination of presumptive endodermal blastomeres in *Xenopus* embryos. *Curr. Biol.* **9**, 869–879 (1999).
- Osada, S. I. & Wright, C. V. E. *Xenopus* nodal-related signaling is essential for mesendodermal patterning during early embryogenesis. *Development* **126**, 3229–3240 (1999).
- Sun, B. I. *et al.* derriere: a TGF-beta family member required for posterior development in *Xenopus*. *Development* **126**, 1467–1482 (1999).
- Zhang, J. *et al.* The role of maternal VegT in establishing the primary germ layers in *Xenopus* embryos. *Cell* **94**, 515–524 (1998).
- Kofron, M. *et al.* Mesoderm induction in *Xenopus* is a zygotic event regulated by maternal VegT via TGF β growth factors. *Development* **126**, 5759–5770 (1999).
- Massague, J. & Chen, Y.-G. Controlling TGF- β signaling. *Genes Dev.* **14**, 627–644 (2000).
- Cheng, A. M., Thisse, B., Thisse, C. & Wright, C. V. The lefty-related factor Xatv acts as a feedback inhibitor of Nodal signaling in mesoderm induction and L-R axis development in *Xenopus*. *Development* **127**, 1049–1061 (2000).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank S. Piccolo for the Cer-S construct, and A. Ashworth, T. Enver, B. Feldman, A. Swain and members of the Jones laboratory for comments on the manuscript.

Correspondence and requests for materials should be addressed to C.M.J. (e-mail: jonesm@icr.ac.uk).

HIV-1 Nef inhibits ASK1-dependent death signalling providing a potential mechanism for protecting the infected host cell

Romas Gelezianas*†, Weiduan Xu*†‡, Kohsuke Takeda§, Hidenori Ichijo§ & Warner C. Greene*||

* Gladstone Institute of Virology and Immunology, PO Box 419100, San Francisco, California 94141-9100, USA

§ Laboratory of Cell Signaling, Graduate School, Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo-ku, Tokyo 113-8549, Japan and CREST, Japan Science and Technology, 5-3, 4-Bancho, Chiyoda-ku, Tokyo 102-0081, Japan

|| Departments of Medicine, and Microbiology and Immunology, University of California, San Francisco, California 94141, USA

† These authors contributed equally to this work

***In vivo* infection of lymphatic tissues by the human immunodeficiency virus type 1 (HIV-1) leads to enhanced apoptosis, which prominently involves uninfected bystander cells^{1–3}. Increased killing of such bystander cells is mediated in part through Nef induction of Fas ligand (FasL) expression^{4–6} on the surface of the virally infected T cells. The subsequent interaction of FasL with Fas (CD95) displayed on neighbouring cells, including HIV-1-specific cytotoxic T lymphocytes, may lead to bystander cell killing and thus forms an important mechanism of immune evasion. As HIV-1 also enhances Fas expression on virally infected cells^{7–9}, it is unclear how these hosts avoid rapid cell-autonomous apoptosis mediated through *cis* ligation of Fas by FasL. Here we show that HIV-1 Nef associates with and inhibits apoptosis signal-regulating kinase 1 (ASK1), a serine/threonine kinase that forms a**

‡ Present address: Rigel Inc., 240 East Grand Avenue, South San Francisco, California 94080, USA.

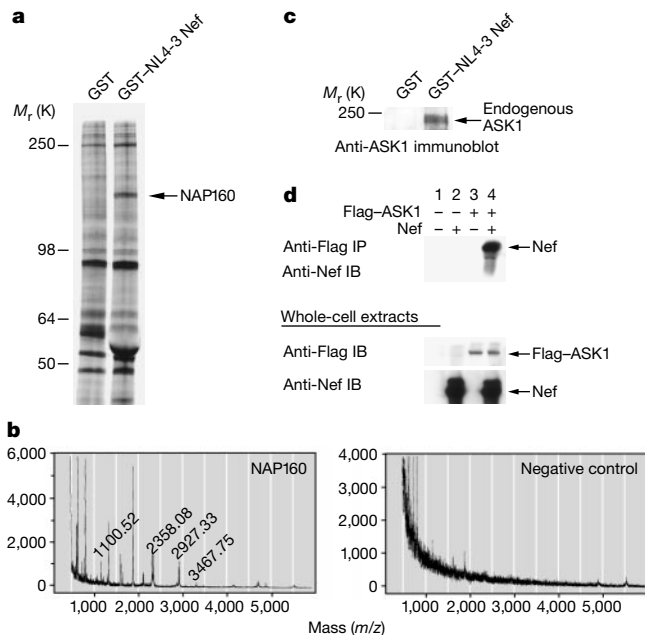


Figure 1 HIV-1 Nef interacts with the serine/threonine kinase ASK1. **a**, A cellular protein of about 160K (NAP160) was specifically captured by GST–Nef, but not GST, matrices. **b**, MALDI–TOF profile of peptides obtained after LysC digestion of unlabelled NAP160 (left) or a region of the same gel lacking identifiable protein (right). **c**, Cellular proteins from Jurkat T-cell lysates captured by GST and GST–Nef matrices were immunoblotted using an anti-ASK1-specific antiserum¹⁰. **d**, 293 cells were transfected with equal amounts of vectors encoding Nef and/or Flag epitope-tagged ASK1. Bottom, whole-cell extracts were immunoblotted with anti-Flag or anti-Nef antibodies to reveal the expression level of each protein. Top, cell lysates were immunoprecipitated with anti-Flag antibodies and subsequently immunoblotted with anti-Nef antibodies to evaluate the interaction of Flag–ASK1 and Nef.

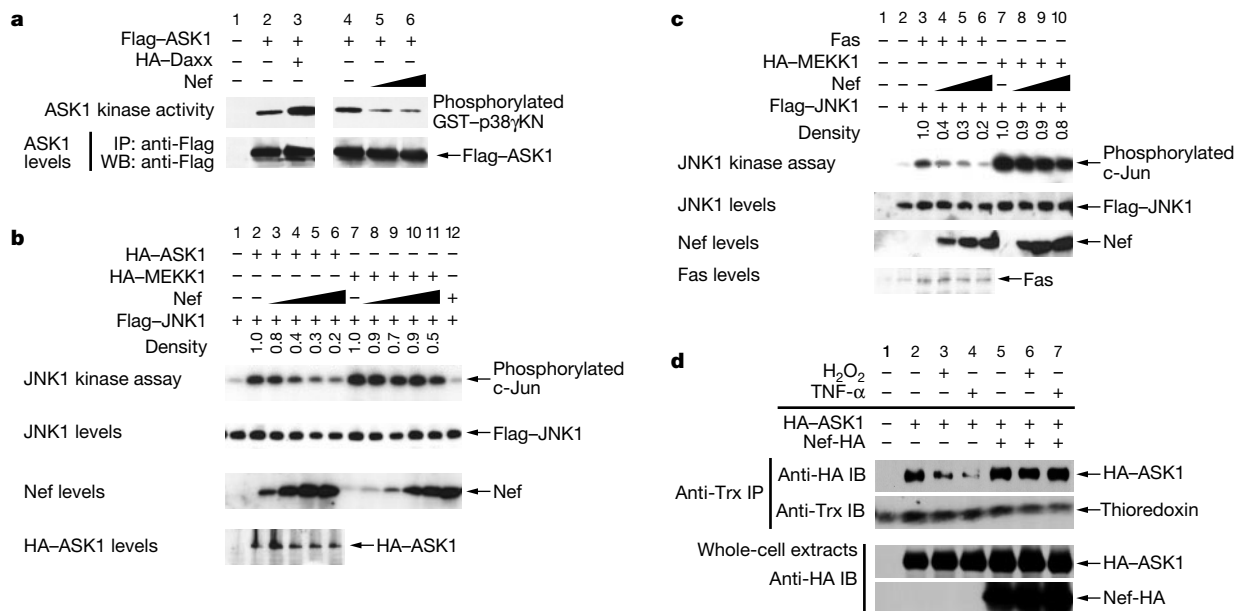


Figure 2 HIV-1 Nef inhibits the intrinsic catalytic activity of ASK1 by preventing thioredoxin release and impairs ASK1- and Fas-mediated induction of JNK1. **a**, 293 cells were transfected with expression vectors encoding HA epitope-tagged Daxx (2 μ g) or Nef (0.5 and 2 μ g) in combination with Flag–ASK1 (0.3 μ g). Flag–ASK1 was immunoprecipitated and subjected to an *in vitro* kinase assay. **b**, Increasing doses of Nef expression plasmid (1, 2, 3.5 and 4.5 μ g) were co-transfected with a plasmid encoding Flag–JNK1 (0.2 μ g) and either HA–ASK1 (0.3 μ g) or HA–MEKK1 (0.3 μ g) into 293 cells. Flag–JNK1 was immunoprecipitated from cell lysates and subjected to an *in vitro* kinase assay in the presence of recombinant GST–c-Jun(1–79). The relative levels of c-Jun phosphorylation

common and key signalling intermediate in the Fas and tumour-necrosis factor- α (TNF α) death-signalling pathways^{10–12}. The interaction of Nef with ASK1 inhibits both Fas- and TNF α -mediated apoptosis, as well as the activation of the downstream c-Jun amino-terminal kinase. Our findings reveal a strategy by which HIV-1 Nef promotes the killing of bystander cells through the induction of FasL, while simultaneously protecting the HIV-1-infected host cell from these same pro-apoptotic signals through its interference with ASK1 function.

Our studies began with a search for cellular proteins that specifically interact with HIV-1 Nef. Using glutathione *S*-transferase (GST)–Nef chromatography and lysates from [³⁵S]methionine and [³⁵S]cysteine radiolabelled Jurkat T cells, we identified a protein with a relative molecular mass (M_r) of about 160,000 (160K) that consistently bound to the GST–Nef matrix but not to GST alone (Fig. 1a). This protein, provisionally designated Nef-associating protein of 160K (NAP160), was digested with the endoprotease LysC, and the resulting peptides were subjected to matrix-assisted laser desorption ionization-time of flight (MALDI–TOF) mass spectrometry (Fig. 1b). Database searches indicated that four peptides in this mixture (1100.52, 2358.08, 2927.33 and 3467.75) corresponded in size to the predicted LysC fragments of ASK1, a member of the mitogen-activated protein kinase kinase kinase (MAP3K) family of serine/threonine kinases¹¹. Immunoblotting of the cellular proteins retained by the GST–Nef matrix with an ASK1-specific antibody confirmed that NAP160 is ASK1 (Fig. 1c). HIV-1 Nef also co-immunoprecipitated with Flag–ASK1 from lysates of dually transfected 293 cells (Fig. 2d, lane 4) but not from lysates lacking Flag–ASK1 (Fig. 2d, lane 2). In contrast, MAPK/extracellular signal-regulating kinase (ERK) kinase kinase 3 (MEKK3), a related MAP3K, did not associate with Nef under these same conditions (data not shown). These findings show that HIV-1 Nef can physically assemble with ASK1 in transfected mammalian cells.

were quantified by laser densitometry. **c**, Flag–JNK1 was co-transfected into 293 cells with a plasmid encoding human Fas or HA–MEKK1 in the presence of increasing doses of a Nef expression vector. Flag–JNK1 was immunoprecipitated and subjected to an *in vitro* kinase assay. **d**, 293 cells transfected with HA–ASK1 and Nef-HA or control vector DNA were treated with H₂O₂ or TNF- α . Endogenous thioredoxin was immunoprecipitated and immunoblotted with anti-HA or anti-Trx antibody (top two panels). Whole-cell lysates were immunoblotted with anti-HA antibody to reveal levels of expression of ASK1 and Nef (bottom two panels).

Several studies have described the involvement of both FasL and TNF- α in HIV-1-induced killing of bystander cells^{4,5,9,13–16}. In addition to Nef-mediated induction of FasL on infected cells^{4–6}, exposure of macrophages to CXCR4-tropic HIV-1 envelope proteins promotes the display of membrane-bound TNF- α that interacts with TNF- α receptors on bystander CD8⁺ T cells, again leading to their apoptosis¹⁴. ASK1 appears to participate as a key signalling intermediate in both the Fas and TNF- α pathways leading to the induction of apoptosis and activation of c-Jun N-terminal kinase (JNK)^{10–12}.

In contrast to Daxx¹⁰, a described activator of ASK1 (Fig. 2a, lane 3), the assembly of Nef and ASK1 inhibited the catalytic activity of this kinase (Fig. 2a, lanes 5 and 6). In terms of the downstream signalling targets of ASK1, Nef produced dose-related inhibition of both ASK1- and Fas-induced activation of JNK1 (Fig. 2b and c, lanes 3–6). Conversely, Nef exerted only weak effects at high dose on JNK1 activation induced by MEKK1, a closely related MAP3K (Fig. 2b, lanes 7–11; Fig. 2c, lanes 7–10). Nef similarly inhibited TNF- α -mediated activation of JNK (data not shown). These inhibitory effects of Nef did not stem from diminished expression of ASK1, JNK1 or Fas, as each remained constant at all doses of the Nef plasmid examined (Fig. 2b and c, bottom panels). Thus, consistent with its ability to inhibit the function of ASK1, Nef suppressed two different death receptor-mediated signalling pathways

(Fas and tumour necrosis factor receptor type I) that commonly use ASK1 to activate JNK^{10–12}.

We next explored the biochemical basis for Nef-mediated inhibition of ASK1 enzymatic activity. Under basal conditions ASK1 assembles with thioredoxin, a cellular factor that inhibits its catalytic function¹⁷. TNF- α -mediated activation of ASK1 requires both the dissociation of thioredoxin from the N terminus of ASK1 and the binding of TNF-receptor-associated factor 2 (TRAF2) to the N- and C-terminal regions of ASK1 (refs 12, 17, 18). Treating 293 cells with either TNF- α or H₂O₂, another ASK1-activating stimulus, induced the dissociation of ASK1 and thioredoxin (Fig. 2d, lanes 3 and 4); however, co-expression of HIV-1 Nef markedly impaired ASK1–thioredoxin dissociation in the presence of these agonists (Fig. 2d, lanes 6 and 7). These findings suggest that Nef inhibits ASK1 catalytic activity by impairing stimulus-coupled release of its normal negative regulator, thioredoxin.

As catalytically inactive versions of ASK1 markedly inhibit both Fas- and TNF- α -mediated induction of apoptosis^{10,11} (Fig. 3b), we next investigated whether Nef negatively regulates Fas- and TNF- α -induced apoptosis. When Jurkat T cells were transfected with either of two HIV-1 wild-type Nef alleles (NL4-3 or SF2) and stimulated for 18 h with TNF- α , we observed that both Nef proteins significantly suppressed TNF- α -induced apoptosis (Fig. 3a). This suppression effect was lost by disrupting the N-terminal myristylation

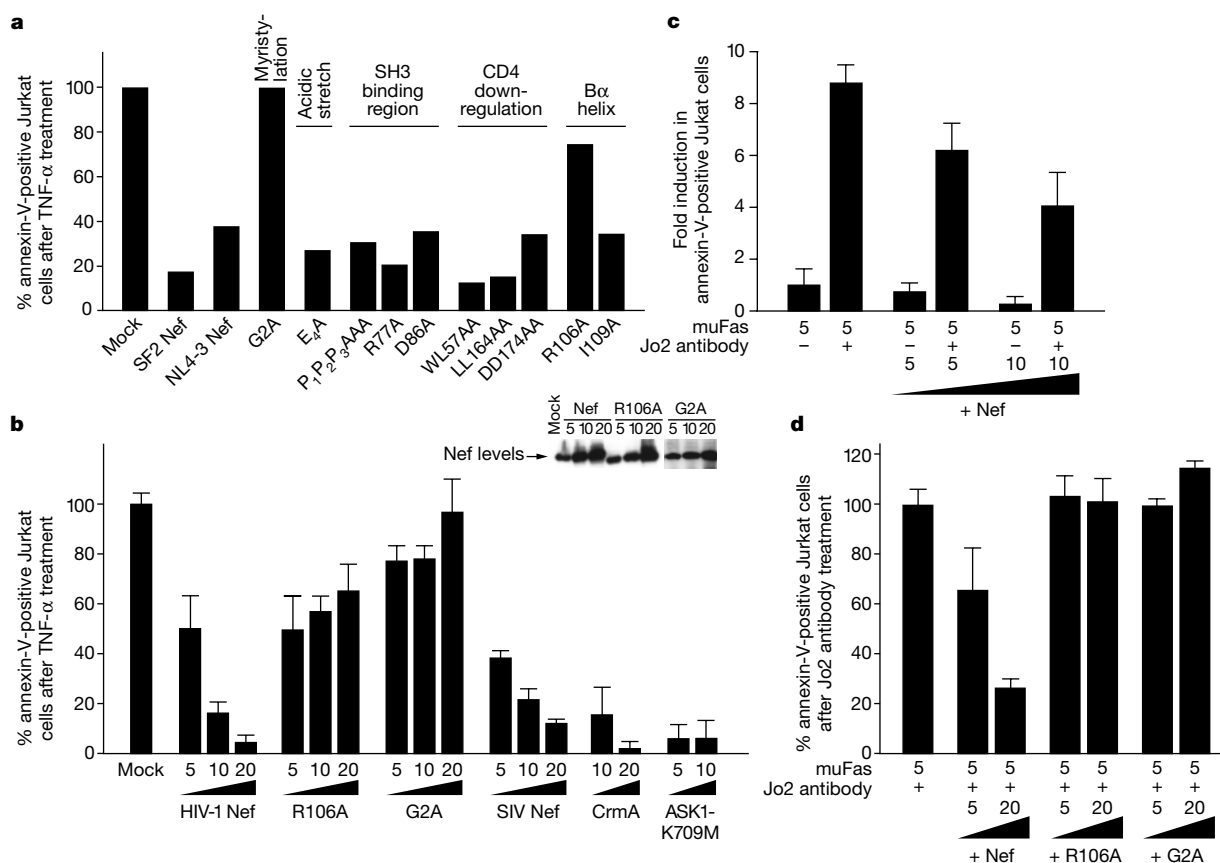


Figure 3 HIV-1 Nef inhibits TNF- α - and Fas-induced apoptosis. **a**, Jurkat cells (1×10^7) were electroporated with expression plasmids encoding either the SF2 or NL4-3 alleles of Nef or NL4-3-derived Nef mutants. After TNF- α treatment, the cells were stained with annexin-V to identify apoptotic cells. Mock-transfected and TNF- α -treated cells were used as controls, and results are expressed as a percentage of this value. **b**, Jurkat cells were electroporated with 5, 10 or 20 μ g of expression plasmids encoding wild-type NL4-3 Nef, Nef R106A, Nef G2A or SIVmac239 Nef or the cowpox virus CrmA protein (10 or 20 μ g) or a kinase-deficient mutant of ASK1 (K709M, 5 or 10 μ g). The transfected cells were treated with TNF- α (100 ng ml⁻¹) for an additional 18 h and stained with annexin-V.

Insert, levels of HIV-1 Nef, R106A and G2A Nef mutants expressed in these lysates. **c**, Jurkat T cells (1×10^7) were electroporated with a murine Fas expression vector (5 μ g) and either 5 or 10 μ g of an HIV-1 Nef plasmid DNA. Cells were either left untreated or stimulated with the anti-murine Fas antibody Jo2. The cells were then stained with annexin-V to assess apoptosis. **d**, Jurkat cells were electroporated with expression vectors encoding murine Fas (5 μ g) and NL4-3 wild-type Nef or the Nef R106A and Nef G2A mutants (5 and 20 μ g). Cells were treated with Jo2 antibody. Results are expressed as the percentage of annexin-V-positive cells relative to murine Fas-transfected and Jo2-treated control cells.

sequence in Nef (G2A) required for its membrane association, or by alanine substitution of R106, implicated in the binding of a member of the p21-activated kinase family¹⁹. Conversely, mutations in the conserved Src homology-3 binding or acidic regions of Nef involved in MHC class I downregulation^{20–22} domains involved in CD4 downregulation^{21–25}, or a second mutation in the B α helix near R106, did not alter the ability of Nef to inhibit TNF- α -mediated apoptosis (Fig. 3a).

These inhibitory effects of wild-type HIV-1 Nef occurred in a dose-related manner (Fig. 3b). However, comparable expression of either the R106A or G2A mutants of NL4-3 Nef (Fig. 3b, inset) did not significantly impair the response. Similarly, the SIVmac 239 Nef allele, which has a key role in the clinical progression of experimentally infected rhesus macaques to AIDS²⁶, produced similar dose-dependent inhibitory effects on TNF- α -induced apoptosis (Fig. 3b). Transfection of these cells with a catalytically inactive form of ASK1 (K709M) generated inhibitory effects comparable to those observed with the cowpox-virus-encoded CrmA protein, a known potent inhibitor of apoptosis (Fig. 3b).

To examine whether Nef similarly suppresses Fas-induced programmed cell death, we transfected a murine Fas expression vector into human Jurkat T cells. Addition of the anti-murine Fas antibody Jo2 selectively crosslinks murine Fas and initiates death signalling in these transfected cells (Fig. 3c). Co-transfection of wild-type NL4-3 Nef produced dose-related inhibition of Jo2-induced cell death (Fig. 3c). As observed previously with TNF- α , neither the G2A nor the R106A mutant of Nef inhibited Jo2-mediated apoptosis in

Jurkat T cells (Fig. 3d). Together, these findings indicate that HIV-1 Nef suppresses apoptosis in human Jurkat T cells triggered by two different death receptors (Fas and TNFR1) that seem to share ASK1 as a common signalling intermediate.

As each of these experiments involved the transfection of Nef expression plasmids into Jurkat T cells, it was important to assess whether the levels of Nef achieved during HIV-1 infection are sufficient to produce similar anti-apoptotic effects. Human SupT1 cells (Fig. 4a) and primary human T lymphocytes (Fig. 4b) were infected with either the NL4-3 HIV-1 strain containing wild-type Nef, a Nef frameshift mutant (Xho), or the R106A Nef substitution mutant. At the peak of infection, these SupT1 cells were treated with TNF- α (100 ng ml⁻¹) for 16 h, whereas the primary human T cells were stimulated with anti-Fas antibodies for 4 h. Significantly greater levels of apoptosis measured by annexin-V staining were observed in both the SupT1 and primary T-cell cultures infected with viruses containing the R106A substitution or Xho frameshift mutants of Nef, as compared with wild-type Nef (Fig. 4a, b). These results indicate that the levels of wild-type NL4-3 Nef achieved during the infection of either of these T-cell targets is sufficient to significantly attenuate TNF- α - and anti-Fas-induced apoptosis.

We next investigated whether Nef assembly with and inhibition of ASK1 catalytic activity was detectable in the infected SupT1 cultures. At the peak of infection, both wild-type Nef and the R106A Nef mutant co-immunoprecipitated with endogenous ASK1 (Fig. 4c). These results confirm the effective interaction of Nef and ASK1 in infected living cells under conditions in which neither protein is overexpressed. Notably, the lack of anti-apoptotic effects by Nef R106A does not derive from its failure to assemble with ASK1. Catalytic activity of endogenous ASK1 was next assessed by immunoblotting lysates from the virally infected and TNF- α -stimulated SupT1 cell cultures with a phospho-ASK1-specific antibody. This antibody reacts with ASK1 only when the enzyme is active (Fig. 4d). Cells infected with HIV-1 NL4-3, expressing wild-type Nef, displayed diminished phospho-ASK1 antibody binding compared with cultures infected with HIV-1 Xho or HIV-1 Nef R106A. Immunoblotting with an anti-actin antibody confirmed comparable protein loading in the three lanes. These findings indicate that the levels of wild-type Nef achieved during HIV-1 infection are sufficient to inhibit ASK1 catalytic activity. Furthermore, the lack of inhibition of ASK1 catalytic activity after infection of SupT1 cells with viruses containing the Xho and R106A Nef

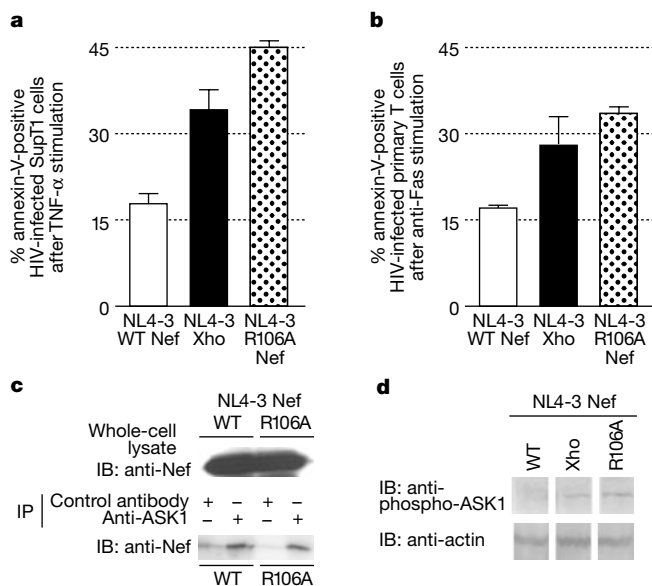


Figure 4 Expression of Nef during HIV-1 infection inhibits TNF- α - and Fas-induced apoptosis. **a**, Human SupT1 cells infected with wild-type HIV-1 (NL4-3), or with isogenic viruses containing a frameshift mutation within Nef (Xho) or Nef R106A were stimulated with TNF- α . Apoptosis was assessed by annexin-V staining. **b**, Human PBMCs infected with NL4-3 wild-type (WT), Xho or R106A Nef were treated with anti-Fas antibody (CH11). The cells were stained with annexin-V, anti-CD3 and anti-p24Gag antibodies. This analysis allowed flow cytometric quantitation of apoptosis occurring in the HIV-1-infected CD3⁺ primary lymphocytes. **c**, Human SupT1 cells were infected with HIV-1 NL4-3 WT or the R106A Nef mutant. When roughly 80% of the cells were infected, cell lysates were immunoblotted with anti-Nef antibodies (top) or immunoprecipitated with either a control antibody or an anti-ASK1 antibody. The immunoprecipitates were analysed by immunoblotting with anti-Nef antibodies (bottom). **d**, After TNF- α stimulation of SupT1 cells infected with NL4-3 WT, Xho or R106A Nef, cell lysates were immunoblotted with antibodies specific for phospho-ASK1, which reacts only with enzymatically active forms of ASK1. Immunoblotting with actin-specific antibodies confirmed comparable protein loading.

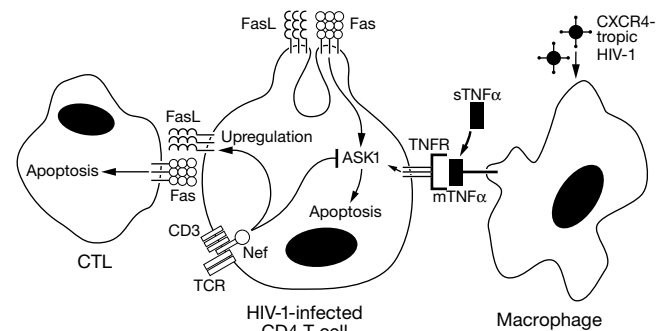


Figure 5 Mechanistic model. HIV-1 Nef causes bystander cell apoptosis through its induction of FasL expression on the infected CD4⁺ T cell. Conversely, Nef inhibits apoptosis in these virally infected host cells through its concomitant suppressive effects on ASK1, a key intermediate in the Fas and TNF- α death signalling cascades. Another proposed mechanism of bystander cell killing involves the induction of membrane-bound TNF- α (mTNF- α) on the surface of macrophages exposed to CXCR4-tropic HIV-1. Membrane-bound TNF- α can, in turn, interact with TNF receptors (TNFR) present on neighbouring cells, delivering a death signal. As TNFR also uses ASK1 as a signalling intermediate, we propose that HIV-1 Nef inhibits soluble TNF- α (sTNF- α)-induced apoptosis in the HIV-1-infected host cell through the same mechanism.

mutations parallels the inability of these viruses to protect their host cells from TNF- α -induced apoptosis (Fig. 4a).

Together, these studies reveal a mechanism by which HIV-1 Nef can alter the intracellular milieu of virally infected host cells by enhancing their resistance to Fas- and TNF- α -induced apoptosis. This action of Nef potentially counters the pro-apoptotic effects accompanying HIV-1 infection mediated through the simultaneous activation of FasL and Fas expression on the infected host cell or the interaction of membrane-bound TNF- α on macrophages with TNF receptors present on the virally infected host cell¹⁴. The inhibitory effect of Nef on ASK1 activity that we describe probably allows these HIV-1-infected host cells to survive long enough to produce new infectious virions (Fig. 5). These findings highlight the ability of Nef to mimic the action of various physiological negative regulators of ASK1, including thioredoxin, 14-3-3 proteins and p21(Cip1/WAF1)^{17,27,28}. In fact, Nef appears to inhibit ASK1 activity by specifically preventing stimulus-coupled release of thioredoxin from ASK1. HIV-1 Nef thus joins a growing list of pathogen-derived proteins such as E2 from the hepatitis C virus and YOPJ from *Yersinia pseudotuberculosis* that effectively target and inhibit cellular kinases, thereby creating an intracellular environment that is more favourable for pathogen survival and spread^{129,30}. □

Methods

GST chromatography

Jurkat T cells (2×10^7) were metabolically radiolabelled with [³⁵S]methionine and [³⁵S]cysteine for 6 h. Lysates were prepared, precleared with glutathione-agarose beads and incubated with either GST or GST–Nef fusion proteins conjugated to glutathione-agarose beads for 2–4 h at 4 °C. We extensively washed the beads in lysis buffer and resuspended them in 2 $\times$ sample buffer. Aliquots were subjected to SDS–PAGE. The gels were dried and exposed to film for autoradiography.

ASK1 and JNK1 *in vitro* kinase assays

We carried out kinase assays essentially as described¹¹. In brief, to assess ASK1 enzymatic activity, transfected ASK1 was immunoprecipitated and subjected to an *in vitro* kinase assay involving the addition of its immediate downstream target MKK3, and a kinase-deficient form of the MKK3 substrate, p38 γ KN, fused to GST. For JNK *in vitro* kinase assays, immunoprecipitated JNK1 was incubated in the presence of the N terminus of c-Jun(1–79) fused to GST. Laser densitometry of the intensity of the *in vitro* phosphorylated c-Jun(1–79) band was performed to quantify JNK1 activity.

Annexin-V staining

Electroporation of Jurkat T cells was performed using a gene pulser (BioRad) (250 V, 950 μ F) incorporating a green fluorescent protein (GFP) expression vector to mark the transfected cells. Eight to eighteen hours after transfection, cells were treated with TNF- α (100 ng ml^{–1} for 18 h) or Jo2 antibodies (24 h) (to crosslink transfected murine Fas). We then stained the cells with the vital dye 7AAD (Pharmingen) and annexin-V conjugated to phycoerythrin (Pharmingen). Cells scoring positive for GFP and annexin-V but negative for 7AAD were analysed by flow cytometry.

ASK1–thioredoxin co-immunoprecipitation assay

We treated transfected 293 cells (5×10^5 cells) with either 5 mM H₂O₂ for 10 min or 200 ng ml^{–1} TNF- α for 15 min. Cell lysates were immunoprecipitated with a goat anti-thioredoxin (Trx) antibody (American Diagnostica) to immunoprecipitate endogenous thioredoxin. These immunoprecipitates were subsequently immunoblotted with either anti-haemagglutinin (HA) or anti-Trx antibodies.

Analysis of apoptosis in HIV-1-infected cells

SupT1 cells or phytohaemagglutinin-stimulated and interleukin-2-treated peripheral blood mononuclear cells (PBMCs) were infected with HIV-1 NL4-3 wild-type or NL4-3 expressing various Nef mutants. We pseudotyped all viruses with VSV-G to enhance the proportion of infected cells. Infections were monitored by measuring intracellular expression of the HIV-1 capsid protein p24Gag. At the peak of infection, when 10–15% of PBMCs (5 d) and about 80% of SupT1 cells (4 d) were expressing p24Gag, the infected SupT1 cultures were treated with TNF- α (100 ng ml^{–1}) for 16 h and the infected PBMCs were treated with anti-Fas antibody (CH11) for 4 h. The cells were stained with Annexin-V-PE and with anti-CD3 antibodies (for PBMCs only), permeabilized, and stained with antibodies specific for p24Gag.

Received 11 October 2000; accepted 26 January 2001.

1. Jaworowski, A. & Crowe, S. M. Does HIV cause depletion of CD4⁺ T cells *in vivo* by the induction of apoptosis? *Immunol. Cell Biol.* **77**, 90–98 (1999).
2. Muro-Cacho, C. A., Pantaleo, G. & Fauci, A. S. Analysis of apoptosis in lymph nodes of HIV-infected

- persons. Intensity of apoptosis correlates with the general state of activation of the lymphoid tissue and not with stage of disease or viral burden. *J. Immunol.* **154**, 5555–5566 (1995).
3. Finkel, T. H. *et al.* Apoptosis occurs predominantly in bystander cells and not in productively infected cells of HIV- and SIV-infected lymph nodes. *Nature Med.* **1**, 129–134 (1995).
 4. Xu, X. N. *et al.* Evasion of cytotoxic T lymphocyte (CTL) responses by nef-dependent induction of Fas ligand (CD95L) expression on simian immunodeficiency virus-infected cells. *J. Exp. Med.* **186**, 7–16 (1997).
 5. Xu, X. N. *et al.* Induction of Fas ligand expression by HIV involves the interaction of Nef with the T cell receptor zeta chain. *J. Exp. Med.* **189**, 1489–1496 (1999).
 6. Hodge, S., Novembre, F. J., Whetter, L., Gelbard, H. A. & Dewhurst, S. Induction of fas ligand expression by an acutely lethal simian immunodeficiency virus, SIVsmmPBj14. *Virology* **252**, 354–363 (1998).
 7. Debatin, K. M. *et al.* High expression of APO-1 (CD95) on T lymphocytes from human immunodeficiency virus-1-infected children. *Blood* **83**, 3101–3103 (1994).
 8. Katsikis, P. D., Wunderlich, E. S., Smith, C. A. & Herzberg, L. A. Fas antigen stimulation induces marked apoptosis of T lymphocytes in human immunodeficiency virus-infected individuals. *J. Exp. Med.* **181**, 2029–2036 (1995).
 9. Sloan, E. M. *et al.* Role of Fas ligand and receptor in the mechanism of T-cell depletion in acquired immunodeficiency syndrome: effect on CD4⁺ lymphocyte depletion and human immunodeficiency virus replication. *Blood* **89**, 1357–1363 (1997).
 10. Chang, H. Y., Nishitoh, H., Yang, X., Ichijo, H. & Baltimore, D. Activation of apoptosis signal-regulating kinase 1 (ASK1) by the adapter protein Daxx. *Science* **281**, 1860–1863 (1998).
 11. Ichijo, H. *et al.* Induction of apoptosis by ASK1, a mammalian MAPKKK that activates SAPK/JNK and p38 signaling pathways. *Science* **275**, 90–94 (1997).
 12. Nishitoh, H. *et al.* ASK1 is essential for JNK/SAPK activation by TRAF2. *Mol. Cell* **2**, 389–395 (1998).
 13. Badley, A. D. *et al.* Macrophage-dependent apoptosis of CD4⁺ T lymphocytes from HIV-infected individuals is mediated by FasL and tumor necrosis factor. *J. Exp. Med.* **185**, 55–64 (1997).
 14. Herbein, G. *et al.* Apoptosis of CD8⁺ T cells is mediated by macrophages through interaction of HIV gp120 with chemokine receptor CXCR4. *Nature* **395**, 189–194 (1998).
 15. Mitra, D., Steiner, M., Lynch, D. H., Staiano-Coico, L. & Laurence, J. HIV-1 upregulates Fas ligand expression in CD4⁺ T cells *in vitro* and *in vivo*: association with Fas-mediated apoptosis and modulation by aurointricarboxylic acid. *Immunology* **87**, 581–585 (1996).
 16. Badley, A. D. *et al.* Upregulation of Fas ligand expression by human immunodeficiency virus in human macrophages mediates apoptosis of uninfected T lymphocytes. *J. Virol.* **70**, 199–206 (1996).
 17. Saitoh, M. *et al.* Mammalian thioredoxin is a direct inhibitor of apoptosis signal-regulating kinase (ASK) 1. *EMBO J.* **17**, 2596–2606 (1998).
 18. Liu, H., Nishitoh, H., Ichijo, H. & Kyriakis, J. M. Activation of apoptosis signal-regulating kinase 1 (ASK1) by tumor necrosis factor receptor-associated factor 2 requires prior dissociation of the ASK1 inhibitor thioredoxin. *Mol. Cell Biol.* **20**, 2198–2208 (2000).
 19. Sawai, E. T., Baur, A. S., Peterlin, B. M., Levy, J. A. & Cheng-Mayer, C. A conserved domain and membrane targeting of Nef from HIV and SIV are required for association with a cellular serine kinase activity. *J. Biol. Chem.* **270**, 15307–15314 (1995).
 20. Lee, C. H., Saksela, K., Mirza, U. A., Chait, B. T. & Kuriyan, J. Crystal structure of the conserved core of HIV-1 Nef complexed with a Src family SH3 domain. *Cell* **85**, 931–942 (1996).
 21. Mangasarian, A., Piguet, V., Wang, J. K., Chen, Y. L. & Trono, D. Nef-induced CD4 and major histocompatibility complex class I (MHC-I) down-regulation are governed by distinct determinants: N-terminal alpha helix and proline repeat of Nef selectively regulate MHC-I trafficking. *J. Virol.* **73**, 1964–1973 (1999).
 22. Greenberg, M. E., Iafraite, A. J. & Skowronski, J. The SH3 domain-binding surface and an acidic motif in HIV-1 Nef regulate trafficking of class I MHC complexes. *EMBO J.* **17**, 2777–2789 (1998).
 23. Bresnahan, P. A. *et al.* A dileucine motif in HIV-1 Nef acts as an internalization signal for CD4 downregulation and binds the AP-1 clathrin adaptor. *Curr. Biol.* **8**, 1235–1238 (1998).
 24. Greenberg, M., DeTulleo, L., Rapoport, I., Skowronski, J. & Kirchhausen, T. A dileucine motif in HIV-1 Nef is essential for sorting into clathrin-coated pits and for downregulation of CD4. *Curr. Biol.* **8**, 1239–1242 (1998).
 25. Craig, H. M., Pandori, M. W. & Guatelli, J. C. Interaction of HIV-1 Nef with the cellular dileucine-based sorting pathway is required for CD4 down-regulation and optimal viral infectivity. *Proc. Natl Acad. Sci. USA* **95**, 11229–11234 (1998).
 26. Kestler, H. W. *et al.* Importance of the nef gene for maintenance of high virus loads and for development of AIDS. *Cell* **65**, 651–662 (1991).
 27. Asada, M. *et al.* Apoptosis inhibitory activity of cytoplasmic p21(Cip1/WAF1) in monocytic differentiation. *EMBO J.* **18**, 1223–1234 (1999).
 28. Zhang, L., Chen, J. & Fu, H. Suppression of apoptosis signal-regulating kinase 1-induced cell death by 14-3-3 proteins. *Proc. Natl Acad. Sci. USA* **96**, 8511–8515 (1999).
 29. Taylor, D. R., Shi, S. T., Romano, P. R., Barber, G. N. & Lai, M. M. Inhibition of the interferon-inducible protein kinase PKR by HCV E2 protein. *Science* **285**, 107–110 (1999).
 30. Orth, K. *et al.* Inhibition of the mitogen-activated protein kinase kinase superfamily by a *Yersinia* effector. *Science* **285**, 1920–1923 (1999).

Acknowledgements

We thank R. C. Wynant of the Stanford University Beckman Center Protein and Nucleic Acid Facility for protein digests and mass spectrometry analysis. We thank E. Verdin, M. Jenkins, V. Linquist-Stepps and T. Liegler for helpful discussions and sharing unpublished data; S. Ferrell, P. Defechereux, P. Bresnahan, L. Gibson, J. Lorens, J. Bogenberger and E. Wieder for technical advice and assistance; J. Carroll, N. Shea, S. Gonzales and C. Goodfellow for preparing figures; S. Ordway and G. Howard for editorial assistance; and R. Givens for preparation of the manuscript. We also thank the following individuals for expression vectors: G. Johnson for MEKK1; M. Karin for Flag-JNK1; M. Jenkins and M. McCune for human Fas; J. Pomerantz and D. Baltimore for murine Fas; and J. Bauer and V. Dixit for CrmA. W.C.G. acknowledges support from the NIH and the UCSF-GIVI Center for AIDS Research.

Correspondence and requests for materials should be addressed to W.C.G. (e-mail: wgreene@gladstone.ucsf.edu).

A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*

Hernando A. del Portillo^{*}, Carmen Fernandez-Becerra^{*}, Sharen Bowman[†], Karen Oliver[†], Martin Preuss[‡], Cecilia P. Sanchez[‡], Nick K. Schneider[‡], Juan M. Villalobos[§], Marie-Adele Rajandream[†], David Harris[†], Luiz H. Pereira da Silva[§], Bart Barrell[†] & Michael Lanzer[‡]

^{*} Departamento de Parasitologia, Instituto de Ciências Biomédicas, Universidade de São Paulo, Av. Lineu Prestes 1374, São Paulo, SP 05508-900, Brazil

[†] Pathogen Sequencing Unit, Sanger Centre, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK

[‡] Hygiene-Institut, Abteilung Parasitologie, Universität Heidelberg, INF 324, D-69120, Heidelberg, Germany

[§] Centro de Pesquisa em Medicina Tropical (CEPEM), Estrada BR 364 – Km 4,5, Porto Velho, Rondônia 78900-000, Brazil

The malarial parasite *Plasmodium vivax* causes disease in humans, including chronic infections and recurrent relapses, but the course of infection is rarely fatal^{1,2}, unlike that caused by *Plasmodium falciparum*. To investigate differences in pathogenicity between *P. vivax* and *P. falciparum*, we have compared the subtelomeric domains in the DNA of these parasites. In *P. falciparum*, subtelomeric domains are conserved and contain ordered arrays of members of multigene families, such as *var*^{3–5}, *rif*^{6,7} and *stevor*⁸, encoding virulence determinants of cytoadhesion and antigenic variation. Here we identify, through the analysis of a continuous 155,711-base-pair sequence of a *P. vivax* chromosome end, a multigene family called *vir*, which is specific to *P. vivax*. The *vir* genes are present at about 600–1,000 copies per haploid genome and encode proteins that are immunovariant in natural infections, indicating that they may have a functional role in

establishing chronic infection through antigenic variation.

P. vivax is the most widely distributed and, except for sub-Saharan Africa, the most prevalent human malarial parasite, causing an estimated 35 million clinical cases annually^{1,2}. Despite its medical importance, molecular analysis of *P. vivax* has been neglected, mainly because parasite material is limited to small quantities obtained from either infected human patients or primates. Continuous *in vitro* culture is hampered by the parasite's host cell specificity for reticulocytes⁹. To initiate studies on the structure and function of the *P. vivax* genome, we have constructed a representative yeast artificial chromosome (YAC) library using parasite material derived from an infected patient¹⁰. The library contains several telomeric YAC clones of which IVD10 was chosen for further investigation. Southern analysis using specific tags derived from IVD10 confirmed that this YAC clone is from *P. vivax* and not a contaminant of human, yeast or *P. falciparum* DNA (data not shown).

IVD10 is 155,711 base pairs (bp) in length. The GGGTT(T/C)A repeats at one end represent consensus plasmodial telomere sequences^{11,12}. Apart from these telomeric repeats, there are no other complex repetitive sequence elements. IVD10 has an overall (A+T) content of 79% (70% in coding and 83% in non-coding sequences), which may not be representative of the entire *P. vivax* genome given that a previous study has reported the presence of (G+C)-rich isochores¹³. IVD10 encodes 32 predicted genes and 4 pseudogenes, yielding a mean gene density of 1 predicted gene every 4.8 kilobases (kb) (Fig. 1), similar to that of *P. falciparum*^{14,15} or *Caenorhabditis elegans*¹⁶.

Most of these (31/32 predicted genes) are members of a new gene family called *vir* (*P. vivax* variant genes). The four pseudogenes present on IVD10 are also related to *vir*, but homology is interrupted by either in-frame termination codons or frameshifts. The remaining gene, *Pvstp1* (*P. vivax* subtelomeric transmembrane protein 1), encoding a putative transmembrane protein of relative molecular mass 98,000 (*M_r* 98K), is unique within IVD10 but has homology to an uncharacterized predicted multigene family in *P. falciparum*, which includes MAL1P3.13 (CAB63568). It also has

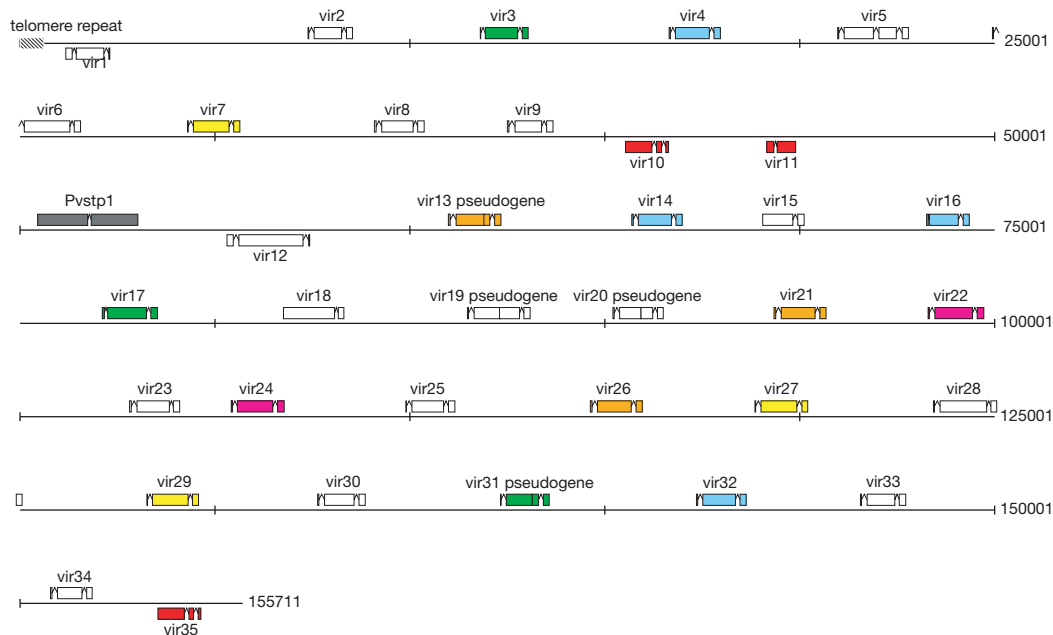


Figure 1 Representation of the *P. vivax* telomeric YAC clone IVD10. The order and orientation of predicted genes and pseudogenes are shown, with exons as boxes and introns as linking lines. *vir* genes and related pseudogenes are consecutively numbered beginning at the telomere repeat sequence (hatched box). *vir* genes that share a high

degree of homology are grouped into subfamilies indicated by colour: subfamily A, green; B, orange; C, blue; D, red; E, fuchsia; F, yellow. The predicted gene *Pvstp1* is shown in grey. The nucleotide numbering starting at the telomere repeats is shown at the end of each line.

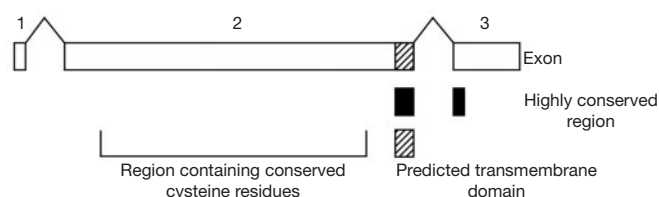


Figure 2 Organization of *P. vivax* *vir* genes: a representation of a consensus *vir* gene showing the 3-exon structure (open boxes) and the predicted transmembrane domain (hatched box). Regions conserved amongst *vir* variants are indicated (filled box).

homology (27.3% identity in 205 amino acids) to the immunovariant SICAVAR antigens in the simian malarial parasite *Plasmodium knowlesi*¹⁷. Sequences, alignments and BLAST analyses can be accessed at http://www.sanger.ac.uk/Projects/P_vivax.

The predicted *vir* genes and pseudogenes have a similar structure. Most have three exons: a short first exon, lacking a consensus signal peptide sequence; a longer second exon containing a predicted transmembrane domain at its 3' end; and a third exon of uniform length (Fig. 2). The region covering the splice boundary between exons 2 and 3 is extremely well conserved and further homologies between family members can be seen in exon 2, with several conserved cysteine residues. The predicted relative molecular masses of VIR proteins range between 26K and 68K, with an average size of 40K. VIR proteins share no homology to known proteins in BLAST analysis. Sixteen of the thirty-two predicted *vir* genes from IVD10 can be grouped into subfamilies on the basis of homology, using CLUSTAL analysis¹⁸ of exons 1 and 2 while excluding the predicted transmembrane sequence. To date, six subfamilies have

been identified, termed A to F.

The *vir* genes are not chromosome specific. They occur on many, possibly all, *P. vivax* chromosomes, as determined by Southern analysis. *P. vivax* chromosomal DNA, isolated from different patients, was hybridized with probes representative of *vir* subfamilies A, B and C (Fig. 3a). However, not all *P. vivax* chromosomes are recognized by a probe to a given subfamily and the hybridization intensities vary. This may suggest that the gene copy numbers differ between the chromosomes or that there is substantial genetic drift within *vir* genes. Notably, *vir* genes are *P. vivax* specific; they do not cross-hybridize to human (not shown), *P. falciparum* or *Saccharomyces cerevisiae* DNA (Fig. 3a). Conversely, *rep20*, a subtelomeric repeat sequence in *P. falciparum*^{14,15}, does not crossreact with *P. vivax* telomeric YAC clones (Fig. 3b). *vir* genes appear to be predominantly located within subtelomeric domains, as shown by Southern analysis of nine telomeric YAC clones derived from different *P. vivax* chromosomes (Fig. 3b). Together, these findings suggest a conserved structural organization of *P. vivax* subtelomeric domains, containing several copies of *vir* genes. There may be 600–1,000 copies present on the 14 *P. vivax* chromosomes, if all *P. vivax* subtelomeric domains are as extensive as IVD10.

vir genes are expressed during intra-erythrocytic parasite development. We isolated *P. vivax* RNA from ten patients, and analysed it by polymerase chain reaction with reverse transcription (RT–PCR) using sets of degenerate oligonucleotides to conserved regions in the *vir* subfamilies A and C. Cloning and sequencing of the resulting RT–PCR products revealed a high degree of variability. Each parasite isolate investigated expressed several *vir* genes, and different sets of expressed *vir* variants were observed in different patients (Fig. 3c). No PCR products were observed in the absence of reverse transcriptase, indicating that the RNA preparations used were free

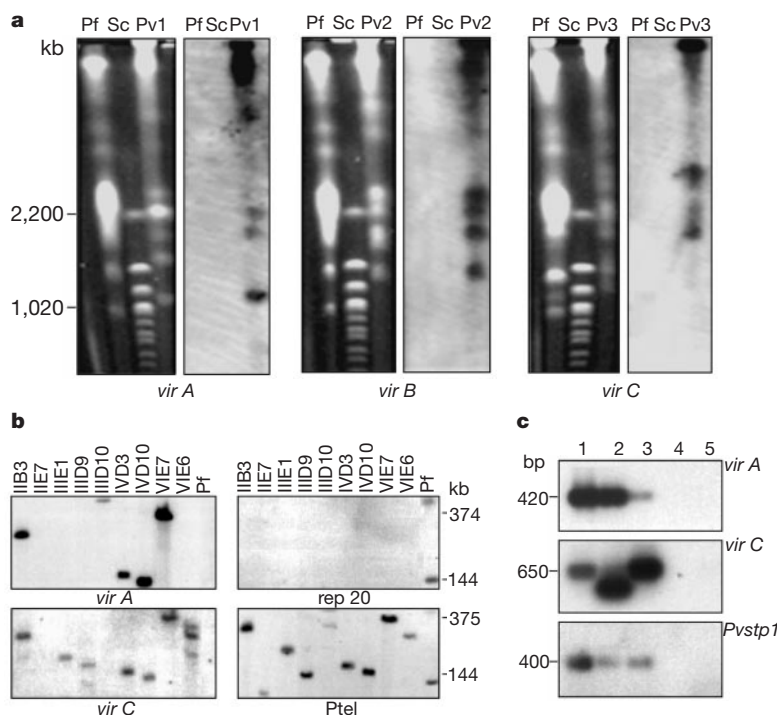


Figure 3 Chromosomal distribution and expression of *vir* genes in *P. vivax*. **a**, Chromosomal DNA from *P. falciparum* (Pf), *Saccharomyces cerevisiae* (Sc) and *P. vivax* (from three different patients, Pv1, Pv2, Pv3) was size fractionated by pulsed-field gel electrophoresis and examined by Southern analysis using probes to the *vir* A, *vir* B and *vir* C subfamilies. Probes were generated by PCR of genomic *P. vivax* DNA from Pv2 using degenerate primers to conserved regions in the respective *vir* subfamily. **b**, Nine random telomeric YAC clones from *P. vivax* and one from *P. falciparum* (Pf) were size fractionated by pulsed-field gel electrophoresis, transferred to nitrocellulose, and hybridized with

specific probes to *vir* A, *vir* C, the *P. falciparum* subtelomeric repeat sequence *rep20*, and an oligonucleotide to the plasmodial telomere repeat sequences (*Ptel*). The *P. vivax* telomeric YAC clones are derived from different chromosomes, as verified by restriction analysis (not shown). **c**, Expression of the *P. vivax* *vir* genes during intra-erythrocytic development. *P. vivax* RNA was isolated from three different patients (lanes 1–3) and analysed by RT–PCR and Southern blot using degenerate oligonucleotides to conserved regions in the *vir* subfamilies A and C, and *Pvstp1* and specific probes, respectively. Human (lane 4) and *P. falciparum* (lane 5) RNA was used as a control.

of genomic DNA contamination. Expression of *vir* variants was confirmed by western analysis of total *P. vivax* asexual blood stage antigens obtained from different *P. vivax* patients. A peptide antiserum to a semi-conserved 15-amino-acid residue motif from the *vir* subfamily D recognized *P. vivax* antigens of around 36K, the predicted average molecular mass of the VIR subfamily D (Fig. 4a). Similar studies revealed that *Pvstp1* is also expressed during intra-erythrocytic development (Fig. 3c).

To determine whether *vir* gene products are immunogenic in natural infections, we expressed the variant exon 2 domains from *vir-4* and *vir-C1-29* as glutathione *S*-transferase (GST) fusion proteins. Fusion proteins were affinity purified and examined in a serological survey using sera from 25 *P. vivax* patients, 5 having a single infection and 20 with histories of multiple (> 3) clinical attacks (Fig. 4b). Only the serum from patient 16 reacted against VIR-4. In the case of VIR-C1-29, only the sera from patients 40 and 41 reacted (Fig. 4b). The three reactive sera were obtained from patients who had experienced several episodes of vivax malaria. In control experiments, all sera recognized total *P. vivax* asexual blood stage antigens, and none of them reacted against affinity-purified GST (data not shown). These results show that VIR proteins are immunogenic in natural infections. Moreover, *P. vivax* patients are apparently exposed to parasites displaying different VIR variants, suggesting that VIR proteins are immunovariant in natural infections. A larger serological study encompassing additional VIR tags

from different subfamilies has validated these conclusions (H.A.d.P. and M.L., unpublished data).

To determine the subcellular localization of VIR proteins, we performed confocal laser scanning immunofluorescence microscopy of purified schizont infected reticulocytes, initially using the affinity-purified anti-VIR-C1-29 antibody fraction from the human serum 40. Purification of VIR-C-specific antibodies was verified in western analyses (data not shown). As expected for antibodies selected for binding to specific variants, only a few parasites reacted (1 per 100). Those that did revealed a strong fluorescence signal at the rim of the host reticulocyte (Fig. 4c). This signal was distinct to that obtained using a mouse antiserum to the *P. vivax* merozoite surface protein 1 (PvMSP-1), recognizing the maturing merozoite within the infected reticulocyte (Fig. 4c). In control experiments, mouse pre-immune serum and human serum from a naive individual showed no reactivity (Fig. 4c). To validate these results, we repeated these studies using the peptide antiserum to the VIR subfamily D. Again, a fluorescence signal at the reticulocyte plasma membrane was observed (data not shown).

Here we have presented the first description of a large variant multigene family in *P. vivax*. Much evidence suggests that *vir* genes encode immunovariant proteins that have a functional role in antigenic variation, and thus in maintaining a chronic infection. Different sets of expressed *vir* tags were detected in parasite isolates obtained from different patients and variant specific immune reactions were observed in natural infections. The apparently high variability in the *vir* family is further reflected by the lack of sterile immunity found even in areas of high *P. vivax* endemicity. VIR proteins were detected at the rim of the host reticulocyte, as opposed to the parasite itself, consistent with a localization of VIR antigens at the surface of the reticulocyte plasma membrane. That *P. vivax* exposes highly polymorphic and immunovariant antigens on its host cell has also been observed¹⁹. It remains to be seen whether these uncharacterized antigens are related to VIR.

vir genes are unique to *P. vivax* and share no homologies to known variant genes in other malarial parasites, including the *P. falciparum* *var*³⁻⁵, *rif*^{6,7} and *stevor*⁸ gene families and the *P. knowlesi* *sicavar* genes¹⁷. Notably, the variant genes identified to date in malarial parasites cluster consistently in subtelomeric domains. Apparently, there is a selective advantage for those genes encoding antigens present on the host cell surface to be located within subtelomeric domains. The high recombination frequencies to which subtelomeric domains are subjected in malarial parasites seem to create an environment that allows the expansion of these genes into gene families and the rapid generation of variants with new antigenic phenotypes²⁰. The species specificity of the subtelomeric domains is in marked contrast to the internal chromosomal domains, which appear highly syntenic between different plasmodial species (H. A. del P., unpublished data). Insight into the evolution of variant genes in malarial parasites, and the function they mediate in antigenic variation and possibly other immune evasion mechanisms, may therefore be gained by a comparative analysis of subtelomeric domains from a broad range of different *Plasmodium* species. □

Methods

Sequencing, assembly, closure and analysis

We sequenced YAC IVD10 as described¹⁵. Briefly, IVD10 was purified by two rounds of pulsed-field gel electrophoresis followed by DNA extraction from the excised band using β -agarase²¹. DNA was sonicated and fragments between 2 and 4 kb were cloned into pUC18. Clones were sequenced using forward and reverse primers and dye terminator chemistry. Contigs were assembled using the PHRAP algorithm (P. Green, unpublished). We filled gaps by a combination of primer walking and sequencing the products of combinatorial PCR. The finished sequence was analysed to identify protein-coding regions, homologies and other features using the Artemis package. This annotation tool allows the output of many searches such as BLAST, fasta and tRNAscan, and homologies to protein families such as Pfam to be viewed and incorporated into the clone annotation. In the absence of other data, many genes on IVD10 were predicted on the basis of homologies to multigene family members within this clone.

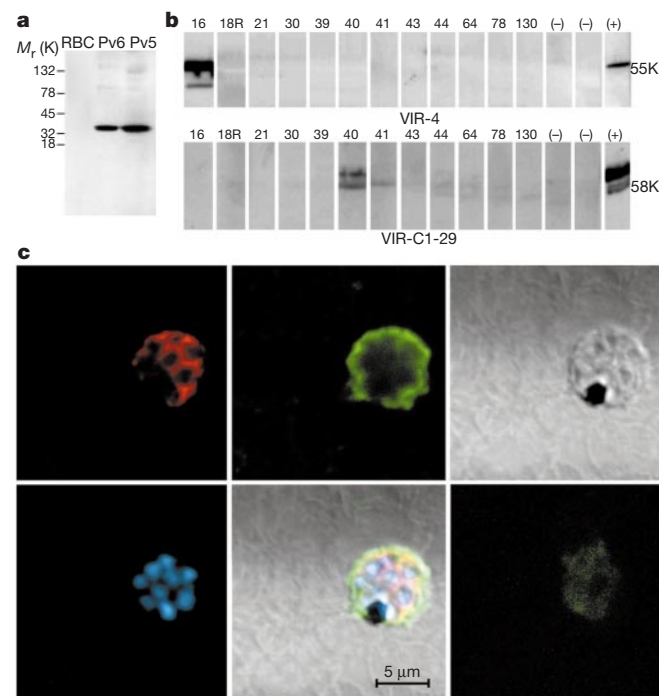


Figure 4 Immunodetection of VIR proteins in *P. vivax*. **a**, Purified total *P. vivax* blood stage antigens derived from two different patients (lanes Pv6 and Pv5) were analysed by western blot using a peptide antiserum to a semi-conserved motif of VIR subfamily D. RBC indicates uninfected erythrocytes. Two micrograms of protein was loaded in each lane. **b**, Serological survey of 12 sera from human patients with histories of multiple (> 3) vivax malaria infections. Sera are indicated by codes. Sera were reacted against affinity-purified VIR-4 and VIR-C1-29 tags expressed as GST fusion proteins. Two naive sera were analysed in parallel (–). A polyclonal anti-GST antiserum was used as a positive control (+). **c**, Confocal laser scanning immunofluorescence of *P. vivax* infected reticulocytes. Enriched *P. vivax* infected reticulocytes were double labelled with antisera to MSP-1 (detected using a Cy3-conjugated secondary antibody) and affinity-purified VIR-C antibodies from patient 40 (detected using a FITC-conjugated secondary antibody). From left to right: anti-MSP-1, human anti-VIR-C, DIC image (top row); Hoechst 33342 stained nuclei, overlay, control using serum from a naive human as the primary antiserum (bottom row).

P. vivax parasite material

We obtained *P. vivax* parasite material from individual adults presenting themselves to the Centro de Pesquisa em Medicina Tropical (CEPEM) in Rondônia, Brazil. These studies received ethical clearance from a local executive committee, and all patients consented to donate their blood.

Pulsed-field electrophoresis and Southern analysis

P. vivax chromosomal DNA was isolated from patients as described¹⁰ and size fractionated by pulsed-field gel electrophoresis using a Bio-Rad CHEF-DRII system (pulsed-field conditions: ramped pulse from 90 to 300 s over 24 h at 5.5 V cm⁻¹ followed by a ramped pulse from 300 to 700 s over 24 h at 4.0 V cm⁻¹ and a final ramped pulse from 700 to 1,200 s over 24 h at 2.5 V cm⁻¹; 1% LE agarose (FMC); running buffer 0.5×TBE at 14 °C. Pulsed-field conditions for telomeric YAC clones have been described¹⁰.

Primers

We used the following degenerate oligonucleotides specific for *vir* subfamilies (5' to 3'): *vir* A, TATTT(A/G)AATT(A/T)TTGGCTA(A/G)A(C/T)(A/G)A(A/G) and (A/G)(C/T)ATTT(A/T/G)GG(A/G)AATGTTTTCAG(C/G)(A/C)ATAT(C/T)AT(A/G); *vir* B, GT(A/T/C)AA(A/T)(A/G)A(A/T)CT(C/T)(C/T)TA(A/G)(A/G)(A/T/G)(A/G)A(A/T)-(A/C)T(C/T)(C/T)A(C/T)AGC(A/C)ATTGA and AGAAGTGTCTCAATAGTGGT; *vir* C, CT(T/G)TT(A/C/G)AA(C/T)TA(C/T)TGGAT(A/G)TA(C/T) and TAC(C/T)(C/T)TATATAA(A/G)GC(A/G)CCAGA(A/T/G)GT(C/G)AT(A/T)G(A/T); *Pvstp1*, CTAGTAACATCATAGAACAC and TTGTTTCCATTCACCTTTAAC.

RT-PCR

We prepared and analysed *P. vivax*, *P. falciparum* and human RNA by RT-PCR as described²². Briefly, RNA isolated using the Tryzol reagent (Gibco-BRL) was treated with DNase in three sequential steps, reaching a final volume of 20 µl. After the last treatment, 1-µl aliquots were amplified by PCR using degenerate primers to *vir* subfamilies A and C and *Pvstp1*. Only RNA preparations that did not amplify any products, and so were free of contaminating genomic *P. vivax* DNA, were used for further analysis. RNA was reverse transcribed using random hexanucleotide primers (Perkin Elmer) and analysed by PCR using degenerate oligonucleotides to *vir* subfamilies. We used annealing temperatures 5 °C below the calculated *T_m* value of the primer pairs. The resulting PCR products were examined by Southern analysis and sequenced.

GST-VIR-C fusion proteins

vir gene tags containing exon 2 were cloned into pGEX-3X (Pharmacia), expressed in *Escherichia coli* and the resulting GST fusion proteins affinity purified, according to the manufacturer's instructions. Purified proteins were examined by western blot analysis and enzyme-linked immunoabsorbent assays, using immune sera from different *P. vivax* patients, as described²³.

Laser confocal scanning microscopy

VIR-C-specific antibodies were affinity purified from the serum of *P. vivax* patient 40 using GST-VIR-C1-29 as the ligand, as described²⁴. The mouse polyclonal anti *P. vivax* MSP-1 has been described²⁵. A guinea-pig antiserum to a semi-conserved amino-acid motif (RKNFKNTYAKKKGLK) in VIR subfamily D was purchased from Eurogentec. We obtained reticulocytes infected with *P. vivax* schizonts from *P. vivax* patients, and fixed them on glass slides as described²⁶. Fluorescein isothiocyanate (FITC)-conjugated affinity-purified F(ab')₂ fragment donkey anti-human IgG + IgM (H+L), Cy3-conjugated affinity-purified donkey anti-mouse IgG (H+L), and FITC-conjugated affinity-purified donkey anti guinea pig IgG (H+L) were purchased from Dianova and used according to the manufacturer's instructions. Hoechst 33342 was purchased from Sigma. Immuno-fluorescence signals were monitored using the LSM 510 (Zeiss).

Data release

The fully annotated sequence of IVD10 (accession number AL360354) is available from http://www.sanger.ac.uk/Projects/P_vivax. The accession numbers for the expressed *vir* gene tags are AF314564 to AF314568, the last corresponding to *vir*-C1-29.

Received 6 November 2000; accepted 27 January 2001.

- David, P. H., del Portillo, H. A. & Mendis, K. N. *Plasmodium vivax* malaria: parasite biology defines potential targets for vaccine development. *Biol. Cell* **64**, 251–260 (1988).
- Galinski, M. R. & Barnwell, J. W. *Plasmodium vivax*: merozoites, invasion of reticulocytes and considerations for malaria vaccine development. *Parasitol. Today* **12**, 20–29 (1996).
- Baruch, D. I. *et al.* Cloning of the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* **82**, 77–87 (1995).
- Smith, J. D. *et al.* Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell* **82**, 101–110 (1995).
- Su, X. *et al.* The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes. *Cell* **82**, 89–100 (1995).
- Kyes, S., Rowe, A. J., Kriek, N. & Newbold, C. I. Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proc. Natl Acad. Sci. USA* **96**, 9333–9338 (1999).
- Fernandez, V., Hommel, M., Chen, Q., Hagblom, P. & Wahlgren, M. Small, clonally variant antigens expressed on the surface of *Plasmodium falciparum*-infected erythrocyte are encoded by the *rif* gene family and are the target of human immune responses. *J. Exp. Med.* **190**, 1393–1403 (1999).

- Cheng, Q. *et al.* *stevor* and *rif* are *Plasmodium falciparum* multicopy gene families which potentially encode variant antigens. *Mol. Biochem. Parasitol.* **97**, 161–176 (1998).
- Golenda, C. F., Li, J. & Rosenberg, R. Continuous *in vitro* propagation of the malaria parasite *Plasmodium vivax*. *Proc. Natl Acad. Sci. USA* **94**, 6786–6791 (1997).
- Camargo, A. A., Fischer, K., Lanzer, M. & del Portillo, H. A. Construction and characterization of a *Plasmodium vivax* genomic library in yeast artificial chromosomes. *Genomics* **42**, 467–473 (1997).
- Ponzi, M., Pace, T., Dore, F. & Frontali, C. Identification of a telomeric DNA sequence in *Plasmodium berghei*. *EMBO J.* **4**, 2991–2995 (1985).
- Vernick, K. D. & McCutchan, T. F. Sequence and structure of a *Plasmodium falciparum* telomere. *Mol. Biochem. Parasitol.* **28**, 85–94 (1988).
- McCutchan, T. F., Dame, J. B., Miller, L. H. & Barnwell, J. Evolutionary relatedness of *Plasmodium* species as determined by the structure of DNA. *Science* **225**, 808–811 (1984).
- Gardner, M. J. *et al.* Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132 (1998).
- Bowman, S. *et al.* The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538 (1999).
- The *C. elegans* sequencing consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012–2018 (1998).
- Al-Khedery, B., Barnwell, J. W. & Galinski, M. R. Antigenic variation in malaria: a 3' genomic alteration associated with the expression of a *P. knowlesi* variant antigen. *Mol. Cell* **3**, 131–141 (1999).
- Persson, B. Bioinformatics in protein analysis. *EXS* **88**, 215–231 (2000).
- Mendis, K. N., Ithamulla, R. I. & David, P. H. Diversity of *Plasmodium vivax*-induced antigens on the surface of infected human erythrocytes. *Am. J. Trop. Med. Hyg.* **38**, 42–46 (1988).
- Freitas-Junior, L. H. *et al.* Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature* **407**, 1018–1022 (2000).
- Vaudin, M. *et al.* The construction and analysis of M13 libraries prepared from YAC DNA. *Nucleic Acids Res.* **23**, 670–674 (1995).
- Kyes, S., Pinches, R. & Newbold, C. A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum*. *Mol. Biochem. Parasitol.* **105**, 311–315 (2000).
- Levitus, G. *et al.* Characterization of naturally acquired human IgG responses against the N-terminal region of the merozoite surface protein 1 of *Plasmodium vivax*. *Am. J. Trop. Med. Hyg.* **51**, 68–76 (1994).
- Hall, R. *et al.* Major surface antigen gene of a human malaria parasite cloned and expressed in bacteria. *Nature* **311**, 379–382 (1984).
- Oliveira, C. L. *et al.* Antigenic properties of the Merozoite Surface Protein 1 gene of *Plasmodium vivax*. *Vaccine* **17**, 2959–2968 (1999).
- Voller, A. & O'Neill, P. O. Immunofluorescence method suitable for large scale application to malaria. *Bull. World Health Organ.* **45**, 524–529 (1971).

Acknowledgements

We thank all the patients who participated in this study; J. d'Arc Neves for collecting blood samples; A. Craig for helping in the initial liaison with the Sanger Centre; M. Quail and the subcloning group; members of team 23 and other members of the Pathogen Unit at the Sanger Centre for their help in sequence generation and analysis; N. Hall for the creation of the web pages; Y. Cully for the graphics; and H. Bujard for encouragement throughout this work. This work was supported by the Deutsche Forschungsgemeinschaft (to M.L.) and the European Commission (to M.L. and H.A.P.).

Correspondence and requests for materials should be addressed to H.A.P. (e-mail: hernando@icb.usp.br).

The ATM–Chk2–Cdc25A checkpoint pathway guards against radioresistant DNA synthesis

Jacob Falck, Niels Mailand, Randi G. Syljuåsen, Jiri Bartek & Jiri Lukas

Institute of Cancer Biology, Danish Cancer Society, Strandboulevarden 49, DK-2100 Copenhagen, Denmark

When exposed to ionizing radiation (IR), eukaryotic cells activate checkpoint pathways to delay the progression of the cell cycle^{1–3}. Defects in the IR-induced S-phase checkpoint cause 'radioresistant DNA synthesis', a phenomenon that has been identified in cancer-prone patients suffering from ataxia-telangiectasia, a disease caused by mutations in the ATM gene^{4–6}. The Cdc25A phosphatase⁷ activates the cyclin-dependent kinase 2 (Cdk2) needed for DNA synthesis^{8,9}, but becomes degraded in response to DNA damage¹⁰ or stalled replication¹¹. Here we report a functional link between ATM, the checkpoint signalling kinase Chk2/Cds1 (Chk2)¹² and Cdc25A, and implicate this mechanism

in controlling the S-phase checkpoint. We show that IR-induced destruction of Cdc25A requires both ATM and the Chk2-mediated phosphorylation of Cdc25A on serine 123. An IR-induced loss of Cdc25A protein prevents dephosphorylation of Cdk2 and leads to a transient blockade of DNA replication. We also show that tumour-associated Chk2 alleles¹³ cannot bind or phosphorylate Cdc25A, and that cells expressing these Chk2 alleles, elevated Cdc25A or a Cdk2 mutant unable to undergo inhibitory phosphorylation (Cdk2AF) fail to inhibit DNA synthesis when irradiated. These results support Chk2 as a candidate tumour suppressor, and identify the ATM–Chk2–Cdc25A–Cdk2 pathway as a genomic integrity checkpoint that prevents radioresistant

DNA synthesis.

To investigate whether degradation of Cdc25A affects the IR-induced S-phase checkpoint, we measured DNA synthesis in human U-2-OS/B3C4 cells engineered to express ectopic haemagglutinin A (HA)-tagged Cdc25A in a tetracycline-repressible manner¹⁰. Endogenous Cdc25A was degraded rapidly in response to IR, along with the dose-dependent inhibition of DNA synthesis (Fig. 1a); however, when the HA–Cdc25A protein was transiently elevated to levels high enough to prevent its degradation, it abrogated the S-phase checkpoint (Fig. 1a), resulting in radioresistant DNA synthesis (RDS) comparable to that seen in cells from ataxia-telangiectasia patients^{4–6}.

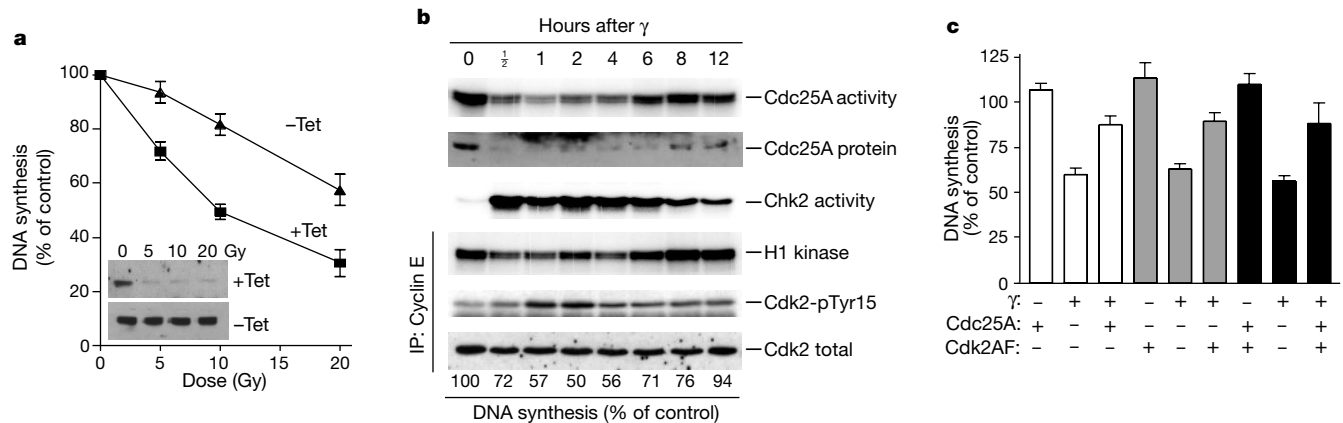


Figure 1 Effects and regulation of IR-induced destruction of Cdc25A. **a**, DNA synthesis and Cdc25A abundance (inset) in γ -irradiated U-2-OS/B3C4 cells repressed (+Tet) or induced (–Tet) 1 h before IR to express ectopic HA–Cdc25A. **b**, Kinetics of IR-induced activities, and abundance of Cdc25A and Chk2, cyclin-E-associated H1 kinase activity

and Cdk2 Tyr 15 phosphorylation in U-2-OS cells exposed to 10 Gy. **c**, DNA synthesis in untreated or 10-Gy γ -irradiated CD20-sorted U-2-OS/T-Rex cells expressing Cdc25A and/or Cdk2AF.

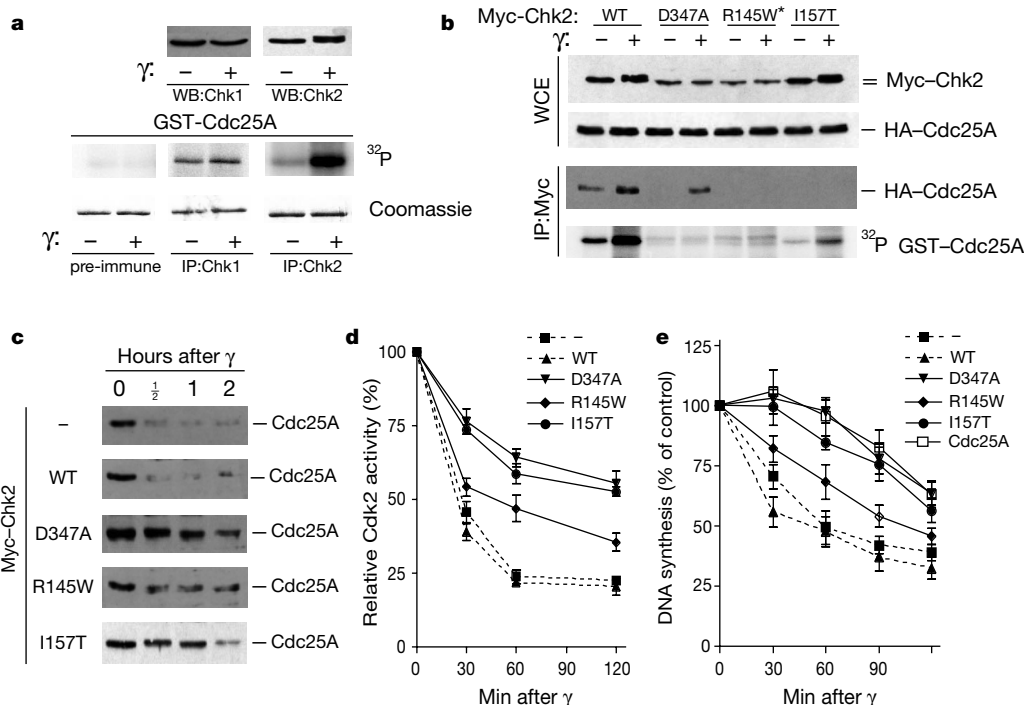


Figure 2 Failure of Chk2 mutants to bind and induce degradation of Cdc25A abrogates the S-phase checkpoint. **a**, Immunoblots of Chk1 and Chk2 and their kinase activities towards GST–Cdc25A in untreated (–) or 10-Gy γ -irradiated (+) U-2-OS cells. **b**, Top, U-2-OS/B3C4 cells were transfected with Myc–Chk2 plasmids as indicated, stimulated to express HA–Cdc25A for 3 h, γ -irradiated (10 Gy; +) or left untreated (–), and immunoblotted for Myc–Chk2 and Cdc25A (WCE, whole-cell extract). Bottom, anti-Myc

immunoprecipitates were assayed for co-precipitated Cdc25A and for *in vitro* kinase activity towards GST–Cdc25A (bottom). Asterisk, a threefold excess of Myc–Chk2-R145W was transfected. **c**, **d**, Various U-2-OS/Myc–Chk2 clones were analysed for Cdc25A abundance (**c**) and Cdk2-associated kinase activity (**d**) at the indicated time points after γ -irradiation (10 Gy). **e**, DNA synthesis in various U-2-OS/Myc–Chk2 clones measured after γ -irradiation (10 Gy). Open squares, U-2-OS/B3C4 cells with induced Cdc25A.

We thought that if destruction of Cdc25A were involved in the S-phase checkpoint, it should be only transient, as the IR-induced inhibition of DNA replication lasts only several hours⁵. Our kinetic measurements showed that Cdc25A was almost completely downregulated by 30 min after irradiation, and recovered to the levels of non-irradiated cells 4–8 h later (Fig. 1b). The maximum downregulation of Cdc25A between 1 and 3 h after IR correlated with an increased inhibitory phosphorylation of Cdk2 on Tyr 15, a reduction of the S-phase-promoting cyclin E/Cdk2 kinase activity, and nearly 50% inhibition of DNA synthesis (Fig. 1b).

These results suggested that the main consequence of the IR-induced Cdc25A downregulation was inhibition of Cdk2, implying that interference with the Cdk2 inhibitory phosphorylation should abrogate the S-phase checkpoint. Indeed, brief conditional expression of the Cdk2AF allele¹⁴, in which the inhibitory Thr 14 and Tyr 15 are replaced by alanine and phenylalanine, respectively, mimicked the effect of Cdc25A overexpression and resulted in RDS (Fig. 1c). Simultaneous expression of both HA–Cdc25A and Cdk2AF did not further increase the degree of RDS (Fig. 1c), indicating that both proteins function in the same pathway.

Ultraviolet-light-induced downregulation of Cdc25A requires the activation of Chk1 (ref. 10), a key signal transducer that, together with Chk2, is implicated in checkpoint pathways activated by damaged or unreplicated DNA^{12,15–19}. To assess whether Chk1 or Chk2 may regulate abundance of Cdc25A on IR, we measured their ability to phosphorylate Cdc25A. The activity and electrophoretic mobility of Chk1 remained unchanged until several hours after IR (Fig. 2a; and data not shown), but Chk2 became activated rapidly, as

judged by its lower electrophoretic mobility and enhanced phosphorylation of GST–Cdc25A 1 h after IR (Fig. 2a). These results suggest that Chk2, but not Chk1, becomes activated with kinetics indicative of a link between IR-induced DNA damage and the rapid inhibition of S phase. Detailed time-course measurements showed that the kinetics of IR-induced Chk2 activity closely paralleled the loss of Cdc25A protein and phosphatase activity, the downregulation of cyclin E/Cdk2 by inhibitory phosphorylation, and the decrease in DNA synthesis (Fig. 1b).

To gain mechanistic insight into this link between Chk2 and Cdc25A, we assessed the IR responsiveness of Myc-tagged forms of human Chk2 transiently transfected into U-2-OS/B3C4 cells, including the engineered catalytically inactive D347A mutant^{12,20}, and the R145W and I157T alleles¹³ with mutations in the putative protein-interaction FHA domain^{12,21,22}, identified in sporadic colon cancer and as a germ-line mutation in the cancer-prone Li–Fraumeni syndrome, respectively¹³. On exposure to IR, the ectopic wild-type Chk2 became rapidly shifted into a more slowly migrating form, and increased its ability to interact physically with transiently elevated HA–Cdc25A *in vivo* and phosphorylate GST–Cdc25A *in vitro* (Fig. 2b). In contrast, basal activities of the three mutants of Chk2 were barely detectable in non-irradiated cells and remained low even after exposure to IR (Fig. 2b).

Notably, the D347A mutant still bound HA–Cdc25A after IR (Fig. 2b), implying that the interaction of Chk2 with its substrate(s) is mediated through its modification by an upstream regulator, independently of Chk2 autocatalytic activity. In contrast, the R145W and I157T cancer-associated mutants of Chk2 could not

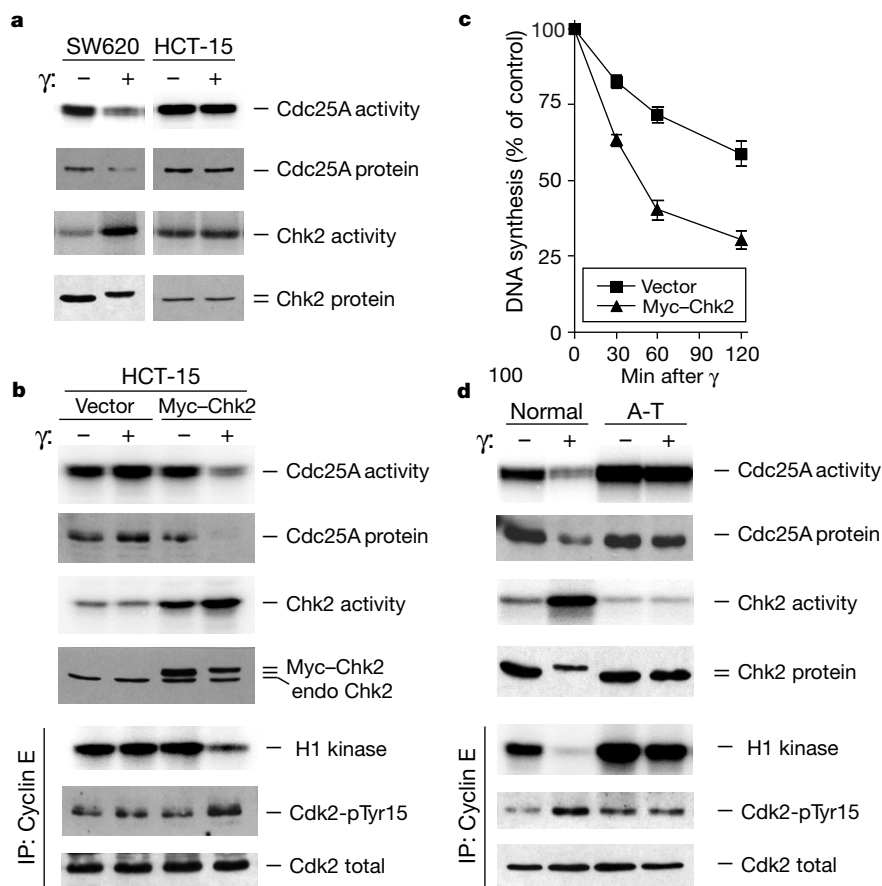


Figure 3 IR-induced destruction of Cdc25A requires functional ATM and Chk2. **a**, Abundance and activities of Cdc25A and Chk2 in untreated (–) or γ-irradiated (10 Gy; +) SW620 and HCT-15 cells. **b**, HCT-15-derived cells were γ-irradiated (10 Gy; +) or left untreated (–) and analysed for activities and abundance of Cdc25A and

Chk2, cyclin-E-associated H1 kinase activity, and Cdk2 levels (Tyr 15 phosphorylated and total). **c**, DNA synthesis in control vector versus Chk2-reconstituted (Myc–Chk2) HCT-15 cells measured after IR (10 Gy). **d**, Normal or ataxia-telangiectasia (A-T) lymphoblasts were treated and analysed as in **b**.

bind HA-Cdc25A, even though the I157T mutant preserved the IR-induced mobility shift (Fig. 2b). Thus, the tumour-associated Chk2 proteins seem defective in their ability to bind and phosphorylate substrates such as Cdc25A.

To examine whether Chk2 and Cdc25A function in a common pathway, we transfected U-2-OS cells with plasmids encoding the Chk2 alleles described in Fig. 2b, together with a gene coding for puromycin resistance. Whereas puromycin-resistant cell populations transfected with wild-type Chk2 or empty plasmid efficiently degraded Cdc25A when exposed to IR (Fig. 2c), cells expressing any of the Chk2 mutants retained substantial amounts of Cdc25A when irradiated (Fig. 2c). Consistent with these differential abilities to modulate Cdc25A protein levels, Cdk2 activity declined in cells transfected with empty vector or wild-type Chk2, but remained elevated in cells expressing the Chk2 mutants (Fig. 2d). Consequently, whereas the control cells responded by inhibiting DNA synthesis, cells expressing the Chk2 mutants failed to impose the S-phase blockade and exhibited RDS comparable to that of cells conditionally overexpressing Cdc25A (Fig. 2e). Thus, Chk2 alleles defective in catalytic activity or interaction with Cdc25A behaved as dominant-negative mutants that abrogated the S-phase checkpoint, which is consistent with Chk2 operating upstream of Cdc25A and Cdk2 in a common pathway.

In addition to Cdc25A, Chk2 also efficiently phosphorylates the p53 tumour suppressor in response to DNA damage^{20,23}; however, the former substrate seems more relevant for S-phase arrest induced by IR, as p53 controls the G1 and sustains the G2/M checkpoints,

rather than the S-phase checkpoint^{24–26}, and RDS occurs independently of p53 status²⁷. Consistent with this, rapid degradation of Cdc25A and loss of its phosphatase activity occurred in SW620 (wild-type for Chk2) but not in HCT-15 cells expressing mutant Chk2 after exposure to IR (ref. 13; Fig. 3a). Notably, both of these colon cancer cell lines express mutant p53 (ref. 13). Re-introduction of wild-type Chk2 into HCT-15 cells to a level exceeding that of the endogenous mutated protein restored the S-phase checkpoint response, including degradation of Cdc25A, increased inhibitory phosphorylation of Cdk2 and downregulation of cyclin E/Cdk2 kinase activity (Fig. 3b), and the ability to arrest S-phase progression (Fig. 3c). These results further support Chk2 as a mediator for rapidly destructing Cdc25A in response to IR, in a p53-independent S-phase checkpoint that protects cells against RDS.

As RDS occurs in cells with a defective ATM gene^{4–6}, as well as through deregulation of Chk2 or Cdc25A, and as ATM activates Chk2 directly (ref. 12), we next compared the response of the Chk2–Cdc25A–Cdk2 pathway in lymphoblasts derived from ataxia-telangiectasia patients to those isolated from normal individuals. Unlike normal lymphoblasts, irradiated ataxia-telangiectasia cells were unable to activate Chk2 and downregulate Cdc25A protein and activity (Fig. 3d). Consequently, exposure of ataxia-telangiectasia lymphoblasts to IR caused neither an increase in Cdk2 Tyr 15 phosphorylation nor an inhibition of cyclin E/Cdk2 kinase activity (Fig. 3d), which is consistent with the well-documented RDS phenotype in these cells⁵.

As signals from damaged DNA are often propagated through

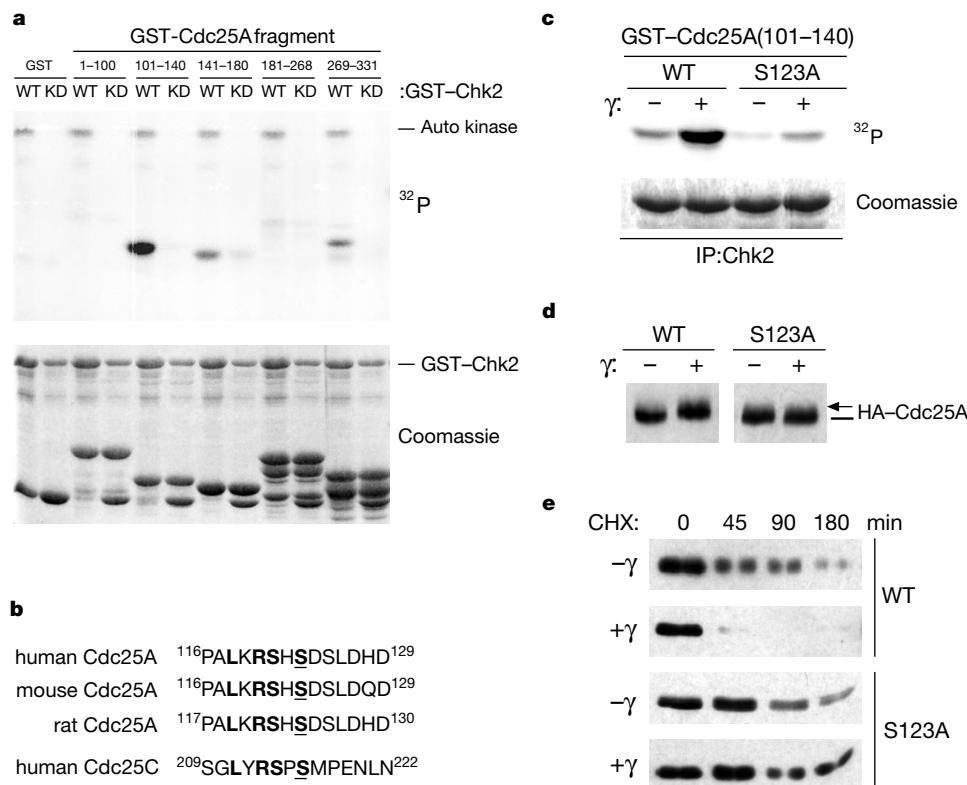


Figure 4 Chk2 phosphorylates Cdc25A on Ser 123 and triggers its IR-induced destruction. **a**, GST–Cdc25A fragments were incubated with purified wild type (WT) or catalytically inactive (KD) GST–Chk2. Proteins resolved by SDS–PAGE were visualized by autoradiography (top) or Coomassie staining (bottom). **b**, Amino-acid sequence flanking Ser 123 in human, mouse and rat Cdc25A, aligned with the Chk1/2-phosphorylated region of human Cdc25C. **c**, GST–Cdc25A(101–140) (WT) or GST–Cdc25A(101–140) (S123A) was incubated with Chk2 immunoprecipitates from untreated (–) or 10-Gy γ -irradiated (+) U-2-OS cells. Proteins were resolved and visualized as in **b**. **d**, U-2-OS

cells transiently transfected with WT or S123A HA–Cdc25A were γ -irradiated (10 Gy; +) or left untreated (–) in the presence of proteasome inhibitor LLnL to prevent destruction, and cell lysates were analysed by immunoblotting. Arrow indicates IR-induced slower-migrating form of Cdc25A. **e**, U-2-OS cells transfected with WT or S123A HA–Cdc25A were γ -irradiated (10 Gy; +) or not (–) 12 h after transfection, incubated with cycloheximide (CHX; 25 μ g ml^{–1}) and immunoblotted for HA–Cdc25A at the indicated times after CHX addition.

phosphorylation cascades, we next identified the Cdc25A residue(s) directly phosphorylated by Chk2, presumably to prime Cdc25A for rapid destruction after IR-induced DNA damage. We made glutathione S-transferase (GST)-coupled fragments derived from the Cdc25A regulatory domain and subjected them to phosphorylation by purified GST-Chk2. The fragment spanning amino acids 101–140 was strongly phosphorylated by wild-type but not by catalytically inactive Chk2 (Fig. 4a), and the sequence flanking Ser 123 of human Cdc25A, conserved in diverse mammalian species, matched the criteria for a Chk2/Chk1 consensus site identified in other proteins such as Cdc25C^{12,16–18} (Fig. 4b). Replacement of Ser 123 with alanine abolished the ability of the IR-activated Chk2 to phosphorylate the corresponding GST–Cdc25A fragment (Fig. 4c). The S123A mutation in the context of full-length Cdc25A resulted in a protein that, unlike wild-type Cdc25A, did not undergo an IR-induced shift when separated on an SDS gel (Fig. 4d), indicating that *in vivo* the S123A substitution abolished an important IR-dependent modification of Cdc25A. Significantly, the S123A mutation rendered Cdc25A resistant to IR-induced degradation under conditions in which a comparable amount of moderately overexpressed wild-type Cdc25A was still effectively degraded *in vivo* (Fig. 4e). We conclude that Chk2-dependent phosphorylation of Cdc25A on Ser 123 represents a critical step in promoting its rapid destruction in response to IR-induced DNA damage.

On the basis of this study, we propose that the ATM–Chk2–Cdc25A–Cdk2 pathway may provide the molecular explanation of the defence mechanism protecting human cells against RDS. The key components and events along this pathway in cells proficient in the S-phase checkpoint are outlined in Fig. 5 (left). The biological and pathophysiological relevance of this mechanism is further supported by the fact that tumour-associated defects of any of its main components cause RDS (Fig. 5, right), and may predispose to, or promote, tumorigenesis (refs 5, 13, 28; and this study). Cells undergoing DNA replication are particularly vulnerable to genotoxic stress including IR; the rapid degradation of Cdc25A and

subsequent silencing of the Cdk2 activity may represent the initial defence barrier, which inhibits DNA synthesis to allow efficient repair. Cdc25A extends the list of important checkpoint mediators targeted by Chk2, currently encompassing the mitotic activator Cdc25C¹², and the p53 (refs 20, 23) and BRCA1 (ref. 29) tumour suppressors, thereby implicating Chk2 in a complex network controlling G1, S and G2/M checkpoints, as well as DNA repair. Such a central role in protecting genome integrity is also consistent with the proposed candidacy of Chk2 for a new tumour suppressor, whose mutations may predispose to tumorigenesis in diverse tissues, as seen in patients with Li–Fraumeni syndrome¹³. Our finding that the tumour-associated Chk2 alleles are indeed loss-of-function mutants provides the missing functional evidence indicating that Chk2 is a genuine tumour suppressor. □

Methods

Antibodies and immunochemistry

Mouse antibody HE172 to cyclin E, as well as immunoblotting, immunoprecipitation, Cdc25A phosphatase activity and *in vitro* kinase assays have been described¹⁰. Rabbit antisera to Cdk2 (M2) and Chk1 (FL-476), and mouse antibody to Cdc25A (F-6, used to detect the endogenous protein) were from Santa Cruz. Rabbit antiserum to Tyr15-phosphorylated Cdk1/Cdk2 was from Calbiochem, and mouse monoclonal antibody to CD20 was from Becton Dickinson. The 9E10 antibody to the Myc epitope was a gift from G. Evan. We generated mouse monoclonal antibodies to human Chk2 (DCS-270) and Cdc25A (DCS-127) by standard hybridoma technology.

Plasmids

To construct the Myc-tagged Chk2 expression vector, the human Chk2 complementary DNA (a gift from S. Elledge) was amplified by PCR with *Pfu* polymerase (Stratagene) and cloned into a pcDNA3 vector (Invitrogen) containing a Myc tag. The catalytically inactive D347A^{12,20}, and tumour-associated R145W¹³ and I157T¹³ mutants of Chk2, as well as the Cdc25A S123A mutant were generated using the QuikChange Site-Directed Mutagenesis kit (Stratagene). We expressed and purified GST–Chk2 and GST–Cdc25A¹⁰ (full length and fragments) according to standard procedures.

Cell culture and RDS assay

We grew U-2-OS cells¹⁰ and SW-620 cells in DMEM containing 10% fetal bovine serum (FBS). HCT-15 cells¹³ and lymphoblasts derived from normal and ataxia-telangiectasia patients (donated by Y. Shiloh) were grown in RPMI with 10% FBS. The U-2-OS-derived B3C4 clone conditionally expressing haemagglutinin (HA)-tagged Cdc25A has been described¹⁰. Expression of the transgene was induced by culturing the cells in tetracycline-free medium for the durations specified in the figure legends. The U-2-OS-derived T-Rex cell line was from Invitrogen. We induced expression of Cdc25A and Cdk2AF by adding tetracycline (1 µg ml^{−1}) 4 h before irradiation (Fig. 1c); ectopic CD20 was expressed throughout the experiment. Isolation of CD20-positive U-2-OS/T-Rex cells with anti-CD20-coupled Dynabeads has been described³⁰. The U-2-OS- and HCT-15-derived polyclonal cell lines stably expressing Myc-tagged wild-type or mutant Chk2 were generated by calcium phosphate transfection (U-2-OS) or electroporation (HCT-15) of cells with empty pcDNA3–Myc vector or pcDNA3–Myc vector containing wild-type or mutant Chk2. A vector encoding puromycin resistance (pBabe-puro) was co-transfected, the cells were selected for neomycin (G418; 400 µg ml^{−1}) and puromycin (1 µg ml^{−1}) resistance, and pooled when visible colonies emerged. Ionizing radiation was delivered by X-ray generator (RT100, Philips Medico; 100 kV, 8 mA, dose-rate 0.92 Gy min^{−1}), and cell extracts were prepared 1 h later except where indicated otherwise. Inhibition of DNA synthesis and RDS after ionizing radiation were monitored as described^{6,26}.

Received 16 November 2000; accepted 11 January 2001.

- Weinert, T. DNA damage and checkpoint pathways: molecular anatomy and interactions with repair. *Cell* **94**, 555–558 (1998).
- Hartwell, L. H. & Kastan, M. B. Cell cycle control and cancer. *Science* **266**, 1821–1828 (1994).
- Elledge, S. J. Cell cycle checkpoints: preventing an identity crisis. *Science* **274**, 1664–1672 (1996).
- Painter, R. B. & Young, B. R. Radiosensitivity in ataxia-telangiectasia: a new explanation. *Proc. Natl Acad. Sci. USA* **77**, 7315–7317 (1980).
- Rotman, G. & Shiloh, Y. ATM: a mediator of multiple responses to genotoxic stress. *Oncogene* **18**, 6135–6144 (1999).
- Stewart, G. S. *et al.* The DNA double-strand break repair gene hMRE11 is mutated in individuals with an ataxia-telangiectasia-like disorder. *Cell* **99**, 577–587 (1999).
- Galaktionov, K. & Beach, D. Specific activation of cdc25 tyrosine phosphatases by B-type cyclins: evidence for multiple roles of mitotic cyclins. *Cell* **67**, 1181–1194 (1991).
- Hoffmann, I., Draetta, G. & Karsenti, E. Activation of the phosphatase activity of human cdc25A by a cdk2-cyclin E dependent phosphorylation at the G1/S transition. *EMBO J.* **13**, 4302–4310 (1994).
- Jinno, S. *et al.* Cdc25A is a novel phosphatase functioning early in the cell cycle. *EMBO J.* **13**, 1549–1556 (1994).
- Mailand, N. *et al.* Rapid destruction of Cdc25A in response to DNA damage. *Science* **288**, 1425–1429 (2000).
- Molinari, M. *et al.* Human Cdc25A inactivation in response to S phase inhibition and its role in preventing premature mitosis. *EMBO Reps* **1**, 71–79 (2000).

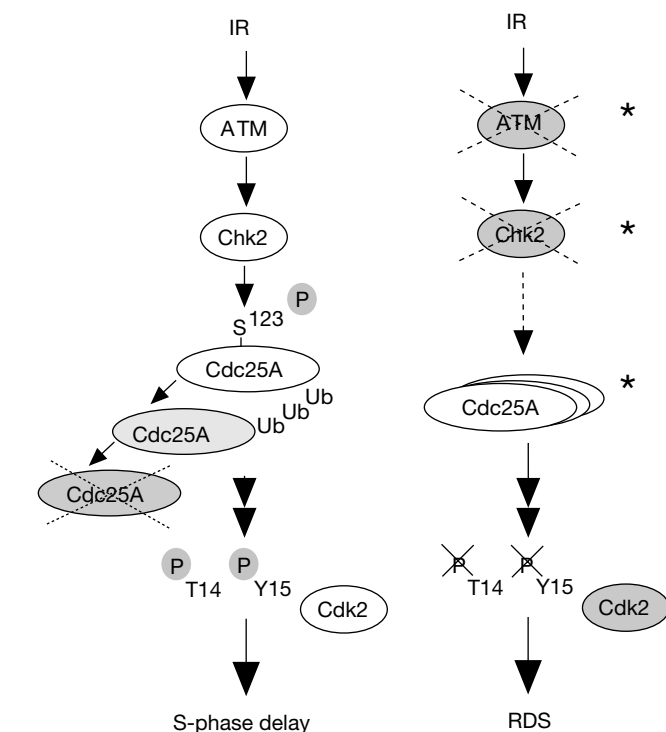


Figure 5 Model of the IR-induced S-phase checkpoint pathway in normal (left) versus checkpoint-deficient (right) cells. Pathway components targeted in cancer are marked by an asterisk. Ub, ubiquitin chains.

12. Matsuoka, S., Huang, M. & Elledge, S. J. Linkage of ATM to cell cycle regulation by the Chk2 protein kinase. *Science* **282**, 1893–1897 (1998).
13. Bell, D. W. *et al.* Heterozygous germ line hCHK2 mutations in Li–Fraumeni syndrome. *Science* **286**, 2528–2531 (1999).
14. Gu, Y., Rosenblatt, J. & Morgan, D. O. Cell cycle regulation of CDK2 activity by phosphorylation of Thr160 and Tyr15. *EMBO J.* **11**, 3995–4005 (1992).
15. Boddy, M. N., Furnari, B., Mondesert, O. & Russell, P. Replication checkpoint enforced by kinases Cds1 and Chk1. *Science* **280**, 909–912 (1998).
16. Furnari, B., Rhind, N. & Russell, P. Cdc25 mitotic inducer targeted by chk1 DNA damage checkpoint kinase. *Science* **277**, 1495–1497 (1997).
17. Peng, C. Y. *et al.* Mitotic and G2 checkpoint control: regulation of 14-3-3 protein binding by phosphorylation of Cdc25C on serine-216. *Science* **277**, 1501–1505 (1997).
18. Sanchez, Y. *et al.* Conservation of the Chk1 checkpoint pathway in mammals: linkage of DNA damage to Cdk regulation through Cdc25. *Science* **277**, 1497–1501 (1997).
19. Zeng, Y. *et al.* Replication checkpoint requires phosphorylation of the phosphatase Cdc25 by Cds1 or Chk1. *Nature* **395**, 507–510 (1998).
20. Chehab, N. H., Malikzay, A., Appel, M. & Halazonetis, T. D. Chk2/hCds1 functions as a DNA damage checkpoint in G₁ by stabilizing p53. *Genes Dev.* **14**, 278–288 (2000).
21. Sun, Z., Hsiao, J., Fay, D. S. & Stern, D. F. Rad 53 FHA domain associated with phosphorylated Rad9 in the DNA damage checkpoint. *Science* **281**, 272–274 (1998).
22. Durocher, D., Henckel, J., Fersht, A. R. & Jackson, S. P. The FHA domain is a modular phosphopeptide recognition motif. *Mol. Cell* **4**, 387–394 (1999).
23. Hirao, A. *et al.* DNA damage-induced activation of p53 by the checkpoint kinase Chk2. *Science* **287**, 1824–1827 (2000).
24. Chan, T. A., Hermeking, H., Lengauer, C., Kinzler, K. W. & Vogelstein, B. 14-3-3 σ is required to prevent mitotic catastrophe after DNA damage. *Nature* **401**, 616–620 (1999).
25. Giaccia, A. J. & Kastan, M. B. The complexity of p53 modulation: emerging patterns from divergent signals. *Genes Dev.* **12**, 2973–2983 (1998).
26. Lim, D. S. *et al.* ATM phosphorylates p95/nbs1 in an S-phase checkpoint pathway. *Nature* **404**, 613–617 (2000).
27. Xie, G. *et al.* Requirements for p53 and ATM gene product in the regulation of G1/S and S phase checkpoints. *Oncogene* **16**, 721–736 (1998).
28. Galaktionov, K. *et al.* CDC25 phosphatases as potential human oncogenes. *Science* **269**, 1575–1577 (1995).
29. Lee, J. S., Collins, K. M., Brown, A. L., Lee, C. H. & Chung, J. H. hCds1-mediated phosphorylation of BRCA1 regulates the DNA damage response. *Nature* **404**, 201–204 (2000).
30. Santoni-Rugiu, E., Falck, J., Mailand, N., Bartek, J. & Lukas, J. Involvement of Myc activity in a G1/S-promoting mechanism parallel to the pRb/E2F pathway. *Mol. Cell. Biol.* **20**, 3497–3509 (2000).

Acknowledgements

We thank S. Elledge, G. Evan, S. I. Reed and Y. Shiloh for providing reagents; K. Hansen for advice; and the Danish Cancer Society, the Human Frontier Science Programme, Alfred Benzon's Fund, the European Commission, the Danish Medical Research Council and the Danish Cancer Research Fund for financial support.

Correspondence and requests for materials should be addressed to J.B. (e-mail: bartek@biobase.dk).

correction

arrow encodes an LDL-receptor-related protein essential for Wingless signalling

Marcel Wehrli, Scott T. Dougan, Kim Caldwell, Louise O'Keefe, Stephanie Schwartz, Dalit Vaziel-Ohayon, Eyal Schejter, Andrew Tomlinson & Stephen DiNardo

Nature **407**, 527–530 (2000).

This paper contained a mistake in Fig. 4a–d. The correct data are described below and are shown on our web site (<http://www.med.upenn.edu/~cellbio/arrow.html>). The main conclusions remain unchanged. In the paper, we used a LacZ enhancer trap inserted at *Distal-less* to investigate its regulation by *arrow*. However, as both *Distal-less* and *arrow* are on chromosome 2R, the LacZ reporter was flipped away in making *arrow* mutant clones. We have repeated these experiments using anti-Distal-less antibodies generating marked, *Minute+ arrow*² mutant clones. Early-induced *arrow* mutant clones do not survive well; those that do survive exhibit loss of Distal-less expression. Later-induced clones show reduced Distal-less, although there is still residual expression. It is possible that the residual expression is due to perdurance of Arrow protein in cells only recently made mutant for *arrow*. As we stated in the paper, these results with *arrow* are quite similar to those obtained for clones mutant for *frizzled frizzled2* (C. M. Chen and G. Struhl, *Development* **126**, 5441–5452; 1999). In the legend to Fig. 4, however, we noted the higher level of Distal-less expression in the wild-type twin spots compared to the heterozygous tissue. Rather than reflecting changes in Wnt signalling, this resulted simply from the presence of two copies of the Distal-less LacZ reporter on the wild-type twin spots. □

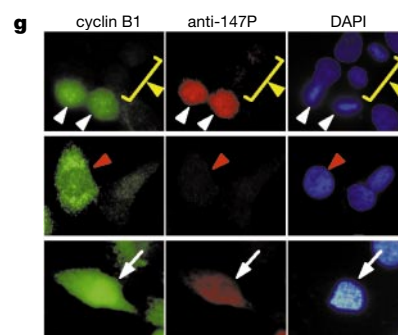
erratum

Polo-like kinase 1 phosphorylates cyclin B1 and targets it to the nucleus during prophase

Fumiko Toyoshima-Morimoti, Eri Taniguchi, Nobuko Shinya, Akihiro Iwamatsu & Eisuke Nishida

Nature **410**, 215–220 (2001).

Figure 5g was printed in black and white instead of in colour. The Figure is reproduced in colour below. □



Contractual dispute

Does the use of fixed-term contracts compromise careers in research?

p849

Have your say

Take part in *Nature's* survey of researchers working under non-permanent contracts

p850

Uncertainty of short-term contracts is turning talent away from science

The proper balance between fixed-term employment and permanent positions has yet to be determined, says Heike Langenberg.

The locksmith who let me into my apartment after I had locked myself out turned out to be a drop-out from a career in academia. When his university adviser had asked him to share his wages (already half-time pay for full-time work) with the lab technician, he decided to look for an alternative career. He thought there must be a better way to earn a living. So he started his own business.

Meagre wages are not the only problem in building a career in research. Many scientists are employed on fixed-term contracts, allowing funding agencies and governments the flexibility to adjust their research agendas.

But one person's flexibility is another's uncertainty. The lack of security and low wages provided by contract research may be turning talent away. The efficiency of research carried out under fixed-term contracts is also in dispute.

Hubert Markl, president of Germany's Max Planck Society (MPS), is content with the situation as it is in his institution. "Contract research is necessary and unavoidable," he says. "If all research was done using permanent positions, science would be paralysed, with no flexibility." More than half of the MPS's scientific staff hold fixed-term positions — allowing them to be employed for a maximum of five years under German law. "The situation is not bad at the moment," says Markl.

Turning point

Dirk Hartung, chairman of the MPS's general works council, a body that represents all the employees, disagrees. "The discussion seems to have reached a turning point," Hartung says. "People are beginning to realize that the excessive use of fixed-term employment affects the quality of research." Hartung believes that continuity is an essential ingredient of good research — as is internal criticism. "People don't dare to dispute their boss's work when they are on a fixed-term contract, for fear of not getting renewal," he points out.

European legislation has directed its



Future shock? Fixed-term contracts provide researchers with flexibility, but they lack security.

attention to fixed-term contracts in several fields, not just science. A directive from the Council of Ministers published in 1999 (1999/70/EC) concerning the framework agreement on fixed-term contracts calls for "achieving the required balance between flexibility and security". The directive must be implemented in national legislation within the European Union (EU) by the end of July 2001. (Ironically, the EU's own research programme is one of the driving forces behind the use of fixed-term contracts in science.)

The new directive aims to limit successive fixed-term employment contracts. It asks member states not to allow people to be hired on a series of fixed-term contracts unless one of three criteria is met. One of these criteria is

a specific reason for the renewal of a fixed-term contract: for example, work on a PhD thesis, which is by its nature limited in time. The second option is to restrict the maximum duration of non-permanent employment — current German law, for example, allows only a maximum of five years of work in science on fixed-term contracts. The third criterion is a restriction on the maximum number of permissible contract renewals.

The directive's impact on scientific research would vary by country. For example, France already has legislation that conforms with the directive, but the United Kingdom and Ireland will have to make significant amendments.

The potential difficulties for researchers

on consecutive fixed-term contracts are manifold. The concern mentioned most often in a survey of the situation of postdocs in Europe, carried out jointly by the European Science Foundation (ESF) and *Nature* (see *Nature* 397, 640–641; 1998), was the uncertainty of contracts and the lack of prospects to obtain permanent positions.

The survey, of people attending ESF conferences who were under 40 and held a PhD, revealed that only a third were in permanent employment. The situation became little better with age and length of experience: half of those aged over 35 were still on fixed-term contracts.

"This research position was intended to be a transitory situation towards a more stable scientific career, although it is becoming a state of permanent uncertainty. There is an anguishing lack of prospect for a career in science for young researchers," writes one survey participant.

France stands out, with three-quarters of the survey participants in permanent positions. France restricts short-term contracts, allowing only a maximum total of 18 months on fixed-term contracts (with the exception of PhD students, who can be employed for up to three years). "Only a very small proportion of researchers are on short-term contracts — most are employed permanently," says Pascal Yiou, a researcher at the Laboratoire de Modélisation du Climat et de l'Environnement near Paris. "From a social point of view, nothing can beat the French system."

But he acknowledges that job security has its downsides. "There is no incentive to go abroad, or to change labs. People tend to stay where they are. This is not very dynamic."

In other countries, contract research is not restricted to young postdocs. Sarah Raper, for example, an internationally recognized climate scientist and lead author in the latest report by the Intergovernmental Panel of Climate Change, works on a fixed-term basis — 23 years after taking her PhD.

"Having failed to get funding in Britain, I now have a three-year contract at the Alfred Wegener Institute in Bremerhaven, Germany. As it is now, I commute," she says. Her family, with two children aged 16 and 14, lives in Norwich, England, where she spent the 19 years until 2000 working on fixed-term contracts funded by various sources, including the US Department of Energy, the



Hubert Markl believes that contract research is 'necessary and unavoidable'.

EU and the UK government. "The flexibility of contract research has suited me," she says, "but it is not a career I would recommend."

Starting a family

The pressures of short-term contracts can increase the difficulties for women trying to combine a family with a career in science. Fixed-term contracts, with no job guarantee after maternity leave and an enormous pressure to produce results within a short time, is not the situation a woman would necessarily choose to be in when she starts a family, unless she is happy to let her career take second place. And women are more likely than men to be employed on fixed-term contracts.

"We cannot accept that women are so far under-represented in science and that they are the first ones to be employed on uncertain contracts," says Achilleos Mitsos, director-general for research at the European Commission. "This is not only unfair and unjust, it is simply a stupid discrimination which can have far-ranging negative effects on the quality of European research."

Speaking for the Max Planck Society, Hubert Markl says: "If you look closer at the situation of women on fixed-term contracts, you find that many young women between about 30 and 36 make a decision about having a family. They often leave science out of choice, much as I find this regrettable, to move where their partners have found a job." He also believes that re-entering a research career, for example after maternity leave, can sometimes even be helped by fixed-term contracts.

Michelle Peckham, now a lecturer at the University of Leeds and a mother of three,

started her family while working on a fellowship from the Royal Society. "At least I had secure funding for five years when I first became pregnant, otherwise I would have been very worried," she says. But it is difficult to find a general pattern that might work. "Trying to fit children into a career structure is very difficult for women," says Peckham. "There are two ways: either you have them very early or very late, after you have become established. But neither of these ways is easy."

Permanently temporary

Sarah Raper points out the difficulties of persuading potential funding agencies to support the research of experienced scientists. "As you get older, it gets harder and harder to get another contract, or a permanent position. You are just not competitive any more. Young postdocs cost half as much," she says. When applying to the EU for project money, just the cost of her wages for the duration of the project will make the work more expensive than most rival proposals. And with limited amounts of money to spend, expensive proposals are less likely to succeed. "I don't know how long I can carry on like this," says Raper. "I have three years in Bremerhaven now, but what will come next, I don't know."

Susanne Rolinski, of the Drittmittler-initiative — a group of activists on contract research — at the University of Hamburg, is employed on a permanent 'soft-money' position. "Research in small chunks of two- to three-year projects wears out researchers through the never-ending duty to write reports for the old project and proposals for new ones," she says. In 2000, the European Commission received 8.3 times more proposals than it could fund, which shows why writing proposals takes up so much of researchers' time.

Colin Bryson, a senior lecturer at Nottingham Trent University whose research interests include fixed-term employment in UK higher education, also questions whether research based on fixed-term positions is cost-effective. "Contract researchers tend to be clustered at the lower end of the pay-scale," he points out. "But we have some evidence that experienced researchers are more than worth their salary." He adds another aspect: "Good research is based on team work. It is very difficult to get a good team together with the high staff turnover that's a consequence of extensive fixed-term contracts."

Where exactly the right balance lies remains debatable. But it seems that this question should be re-investigated from the perspective of efficiency as much as of social acceptability. ■

Heike Langenberg is a physical sciences editor at *Nature*.

European directive

♦ http://europa.eu.int/eur-lex/en/lif/dat/1999/en_399L0070.html

UK consultation document

♦ <http://www.dti.gov.uk/er/fixed/index.htm>

Have your say

We would like to explore the concerns associated with 'soft-money' research on a global basis, and are therefore inviting you to participate in our current survey on the situation of researchers on non-permanent contracts and their need for information.

If you are working as a researcher in a non-tenured position, we would very much appreciate your cooperation. Please assist us by going to the web page <http://www.nature.com/Marketing/ques> and fill in our questionnaire. We will use the information gathered to highlight the concerns of researchers without permanent employment, and to improve *Nature's* service to this group of scientists.

H. L.